Design and Implementation of a Multi-tier Electroencephalography Visual and Emotion Recognition System

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

2019

Mohammed Abdulaal
School of Electrical and Electronic Engineering
Contents

Contents 2
List of figures 7
List of tables 9
List of publications 10
List of prizes 11
Abstract 12
Declaration of originality 13
Copyright statement 14
Acknowledgements 15

1 Introduction 16

1.1 Setting the scene ........................................... 16

1.1.1 The purpose of this study ........................... 16

1.1.2 Electroencephalography (EEG) and brain-computer interfaces (BCIs) .................................................. 16

1.1.3 The scope of the study: Visual and emotion-based BCIs .... 18

1.2 Thesis organisation ........................................ 18

1.2.1 Introduction ........................................... 18

1.2.2 Chapter 3: Performance robustness ....................... 19

1.2.3 Chapter 4: User experience as a measure of performance .... 19

1.2.4 Chapter 5: Quantifying violin quality ....................... 20

1.2.5 Summary ............................................ 20

References ................................................... 21
2 An overview of visual and emotion brain-computer interfaces:

EEG hardware and signal processing

2.1 Introduction .......................................... 23

2.2 EEG recorders ....................................... 28
   2.2.1 Introduction ................................... 28
   2.2.2 Research-grade active electrode EEG recorders: Brain Products ActiChAmp and Biosemi Active Two ................. 29
   2.2.3 Commercial-grade EEG recorder: Emotiv EPOC+ ............ 30
   2.2.4 Summary ...................................... 30

2.3 Feature extraction .................................... 31
   2.3.1 Introduction ................................... 31
   2.3.2 Time-points ................................... 31
   2.3.3 Power spectral density (PSD) ......................... 32
   2.3.4 Covariance matrices ................................ 33
   2.3.5 Statistical features ................................ 34
   2.3.6 Digital wavelet transformation (DWT) .................... 37
   2.3.7 Summary ...................................... 37

2.4 Spatial Filtering ...................................... 37
   2.4.1 Introduction ................................... 37
   2.4.2 Unsupervised ICA ................................ 38
   2.4.3 Supervised CSP .................................. 40
   2.4.4 Supervised filter-bank- and Riemannian-geometry-based CSPs . 42
   2.4.5 Supervised xDAWN ................................ 44
   2.4.6 Summary ...................................... 45

2.5 Classification ....................................... 46
   2.5.1 Introduction ................................... 46
   2.5.2 Logistic regression (LR) .......................... 46
   2.5.3 Support vector machines (SVM) ....................... 47
   2.5.4 Neural networks (NNs) ............................ 48
   2.5.5 Riemannian-geometry-based minimum distance to the mean classification ................................. 51
   2.5.6 Cross-validation methods ......................... 52
   2.5.7 Area under the receiver operating characteristic curve (AUC) . 54
   2.5.8 Statistical significance using the t-test ............... 55
2.5.9 Summary ........................................... 56
2.6 Chapter Summary .................................... 57
References ............................................. 59

3 An evaluation of cross-validation and classifier methods for visual and emotion BCIs: Recommendations for robust performance 67
3.1 Introduction .......................................... 67
3.2 Critical analysis of cross-validation methods and their impact on NN performance inflation ........................................... 68
3.2.1 Introduction .......................................... 68
3.2.2 Motivation ........................................... 68
3.2.3 Methods ............................................ 70
3.2.4 Results ............................................. 74
3.2.5 Discussion ........................................... 77
3.2.6 Summary ............................................ 79
3.3 Demonstrating logistic regression (LR) is the best performing classifier for low-channel count EEG in visual and emotion classification ........................................... 79
3.3.1 Introduction .......................................... 79
3.3.2 Aim .................................................. 80
3.3.3 Methods ............................................ 81
3.3.4 Results ............................................. 89
3.3.5 Discussion ........................................... 92
3.3.6 Summary ............................................ 94
3.4 Chapter summary ..................................... 95
References ............................................. 96

4 Improving user experience performance: A new non-flickering visual BCI paradigm 107
4.1 Introduction .......................................... 107
4.2 The face paradigm: Visual perception-based communication ........................................... 109
4.3 Methods ............................................. 113
4.3.1 Introduction .......................................... 113
4.3.2 User experience survey ........................................... 113
4.3.3 Wirelessly synchronised commercial-grade EEG ........................................... 115
5 A novel use of emotion BCIs: Classification of violins using EEG to determine violin type perception

5.1 Introduction .................................................... 138
5.2 Methods ......................................................... 141
  5.2.1 Introduction ................................................ 141
  5.2.2 EEG listening tests ........................................ 141
  5.2.3 Expressed emotion survey ................................. 143
  5.2.4 Riemannian CSP spatial filtering and LR classification .. 143
  5.2.5 Summary .................................................. 146
5.3 Results and discussions ....................................... 147
  5.3.1 Emotional response survey ............................... 147
  5.3.2 Replication of emotion EEG classification .............. 147
  5.3.3 Violin type classification ............................... 150
5.4 Chapter summary ............................................ 155
References ....................................................... 156

6 Conclusions and future work ................................ 163
  6.1 The purpose of the study .................................... 163
  6.2 Thesis summary and contributions .......................... 164
  6.3 Future directions ............................................. 166
  6.4 Conclusion .................................................. 168
References ....................................................... 168

Appendices ....................................................... 170

A SVM nested and non-nested cross-validation study ....... 171
B Postgraduate poster showcase winner 177

C Violin type commercial-grade EEG pilot study 179
# List of figures

2.1 BCI architecture .................................................. 23
2.2 Typical BCI setup .................................................. 25
2.3 BCI paradigms ....................................................... 26
2.4 Examples of ERPs ................................................... 27
2.5 Spatially filtered foot and hand imagined movement ................. 42
2.6 Overall process of the filter-bank common spatial patterns (FBCSP) ... 43
2.7 xDAWN spatial filtering ............................................ 45
2.8 Representation of hyperplane optimisation .......................... 47
2.9 Representation of deep neural networks ............................ 49
2.10 Representation of the relationship between SPD manifold and tangent space .................................................. 51
2.11 Nested and non-nested cross-validation methods explained .......... 53
2.12 Nested and non-nested cross-validation accuracies on Iris dataset .......... 54
2.13 Presentation of the ROC curve ..................................... 55
2.14 Graphs of p-values ................................................ 56
3.1 A study utilising a non-nested cross-validation method ................. 69
4.1 The communication interface ....................................... 110
4.2 User experience experiment setup .................................. 114
4.3 The setup for the system .......................................... 116
4.4 An overview of the synchronisation device .......................... 117
4.5 Circuit diagram for the synchronisation device ....................... 118
4.6 Synchronisation triggers ............................................ 119
4.7 12 ICA components ............................................... 121
4.8 ICA decision-making process ...................................... 122
4.9 ERPs before and after ICA filtering ................................ 123
4.10 Algorithm for classification of faces and non-faces .................. 125
5.1 EEG listening test experiment setup. ........................................ 142
5.2 Algorithm for classification of emotive states ............................ 144
5.3 Valence and arousal CSPs ..................................................... 149
5.4 The process of pairing the clips. ........................................... 152
5.5 Violin type CSPs ............................................................ 154
List of tables

3.1 Snapshot of emotion recognition studies utilising two different EEG
recorders ................................................................. 71
3.2 Literature survey of studies testing their algorithms on the DEAP
dataset. ............................................................. 74
3.3 Nested and non-nested cross-validated classification of high and low
calence using MLPs and PSD features for the DEAP dataset [12] .... 76
3.4 Snapshot of emotion recognition studies utilising DEAP dataset .... 77
3.5 204 MEG classification accuracies (AUC) and p-values for LR, SVM,
and NN ................................................................. 90
3.6 74 EEG classification accuracies (AUC) and p-values for LR, SVM,
and MLP ............................................................... 90
3.7 12 EEG classification accuracies (AUC) and p-values for LR, SVM,
and MLP ............................................................... 91
3.8 Valence classification accuracies (AUC) and p-values for LR, SVM,
and MLP ............................................................... 93
3.9 Arousal classification accuracies (AUC) and p-values for LR, SVM,
and MLP ............................................................... 94
4.1 List of 50 questions used in the paradigm ............................. 112
4.2 User experience survey results; 50 participants .................... 127
4.3 Face versus non-face classification accuracies; 10 participants .... 128
5.1 Expressed emotion survey results; 30 participants .................. 147
5.2 Valence and arousal classification accuracies; 16 participants .. 148
5.3 Electric versus acoustic violin classification accuracies; 16 participants . 151
5.4 Emotions separated electric versus acoustic violin subject-independent
classification accuracies; 16 participants ............................ 151
5.5 Paired electric and acoustic violin classification accuracies; 16 partici-
pants ................................................................. 153
List of publications


List of prizes

1. The first prize in the School of Electrical and Electronic Engineering post-graduate research poster conference in November 2016. Poster title: *Electroencephalography emotion recognition: Quality of musical instruments based on listener’s brain behaviour*. The poster is enclosed as Appendix B.

2. The first prize in the Manchester Enterprise Centre Venture Further business start-up competition in the business category. The prize was won part of a team of 5 members. The competition involved presenting to a panel of judges that was co-delivered by the author of the thesis.
Abstract

Brain-computer interface (BCI) is a fast growing field of research that utilises non-invasive Electroencephalography (EEG) to enable control and communication for both patients and healthy users. The aim of this thesis is to provide a comprehensive improvement step for BCI technology. This will be focussed around three key values; robustness, human-centred development, and innovation. Firstly, BCI robustness is enhanced by critically analysing the literature and, for the first time, justifying the variations in classification accuracies. This was achieved by conducting a literature survey of emotion recognition studies and concluding, by re-implementation, that the cause of these variations is the use of non-nested cross-validation when testing a neural network (NN) classifier. Secondly, human-centred development is conducted by examining user experience and reducing user fatigue associated with flickering lights in visual BCI. This was achieved by design and implementation of a non-flickering visual BCI paradigm that enables communication by looking at pictures of faces and non-face. The brain behaviour of visual perception was then used classify choices. The convenience of non-flickering picture was validated by 88% of participants in a survey, and the paradigm was tested using a commercial-grade EEG. Finally, innovation is created by a novel application of emotion recognition using EEG to classify violin types associated with different sound qualities. The aim of the experiment was to provide insights into brain activity during the perception process. This was achieved by conducting EEG listening tests of specifically-comprised melodies to express various emotions, played by the two violins. The two violins were classified with a statically significant accuracy of 61%. This thesis provides contributions in various ways, such as literature reviews, classifier optimisation, electronic design and implementation, survey conducting, EEG data collection, and study replications.
Declaration of originality

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright statement

i The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.
Acknowledgements

I would like to express my sincere gratitude to my supervisors Prof. Patrick Gaydecki and Dr. Alex Casson for their continuous support and patience. I would also like to thank my father Prof. Jamal Abdulaal, my mother Eman Hajjar, who have always been my true mentors and friends, my wife Ola Allam, who has been the most patient with me, my son Abdullah for making me forget my stress, and my unborn daughter Eman.

Besides, my sincere thanks goes to my friends and lab colleague: Dr. Eleanor Trimble, Dr. Thomas Lloyd, Dr. Siddharth Kohli, Dr. Keith Philips, Dr. Meera Duleabh, Dr. Tahmina Zebin, Wenyang Xie, Gareth Jones, Erdem Atbas, Nikhil Jacob, Andrius Vabalas, and many others.
Chapter 1

Introduction

1.1 Setting the scene

1.1.1 The purpose of this study

This thesis aims to make a comprehensive improvement in making brain-computer interfaces (BCIs) a better field. Before making any engineering contributions, existing literature is critically analysed in order to establish robustness in BCI tests. As part of this analysis, the performance measures are challenged to produce new reliable understandings. However, in order for this technology to reach the market, robustness is not the only concern, user experience also has to be accounted for. Without this consideration, BCI research is in danger of becoming less appealing to users. This factor is why reducing user fatigue is the second contribution of this thesis. Subsequently, this technology is extended to introduce an innovative emotion recognition system, capable of classifying violin types, to help understand our brain. The thesis focuses on three values; robustness, human-centred development, and innovation. Each value is represented by the work accomplished in Chapters 3-5, respectively.

1.1.2 Electroencephalography (EEG) and brain-computer interfaces (BCIs)

Electroencephalography measures brain activity and is primarily used to assess epilepsy and other brain disorders. As brain disorders are commonly identified by unusual activity in the brain, the frequency of a measured signal often indicates if a patient is healthy or not [1]. The output of an EEG device resembles waves with different spectral frequencies, and the EEG signals which are useful for interpreta-
tion are in the 2-40 Hz frequency range. In order to measure electrical activity in the brain, most EEG use non-invasive electrodes that require no surgery or extensive preparation, which contributes to the current popularity of EEG over other analytical methods. However, while EEG devices are useful, there are many challenges in designing them. The low amplitudes of order $\mu$V EEG signals are one design challenge. Another challenge comes from the surrounding signals, such as electromyography (EMG) signals caused by muscular movement, that can mix with, and degrade, the EEG signals. Typically, EMG signals can be up to 100 times stronger in amplitude than EEG signals.

Brain-computer interfaces (BCIs) utilise EEG to link the brain with external machines for various applications, including medical analysis, environmental control, and communication for patients with partial or total paralysis [2]. Although these patients have lost control over some of their muscles, they may still have the cognitive ability to use a BCI, which can improve their quality of life. BCIs can also help people meditate by monitoring the brain’s activity level [3]. In addition, BCIs are used in both single and multiplayer video games, such as in fighting games that measure high brain activity during a specific period of time and then graphically represent it [4]. In addition to their active use, EEG has been used in other applications such as in neuro-marketing. It can be exploited to analyse and predict consumer behaviour and is often utilised to improve marketing campaigns [5]. Commonly used BCI systems are based on event-related potentials (ERPs), steady state visually evoked potentials (SSVEP), motor imagery (MI), and emotion recognition. Summarily, BCI applications vary extensively and have been successful in many different scientific fields.

Historically, Professor Jacques Vidal, commonly recognised as the founder of BCIs, introduced the term BCI in the 1970s. However, Professor Hans Berger was the first to discover that the brain’s electrical signals could be recorded from the scalp in 1924 [2]. Nowadays, according to Transparency Market Research, the international market size of BCIs is expected to reach $1.2b by 2024 with a compound annual growth rate of 14.0% during 2016-2024 [6]. In 2013, 85% of the total market revenue was for non-invasive BCIs, and this revenue is showing steady growth. In particular, the increases in diagnoses of brain disorders and in government funding for healthcare solutions related to these disorders has led to a strong demand
for BCIs [7].

1.1.3 The scope of the study: Visual and emotion-based BCIs

This thesis focuses on two applications; visual BCIs which utilise screens to enable communication, and emotion-based BCIs which use EEG to predict valence and arousal. The thesis scope includes the design and implementation of a low-level EEG recognition system, which can be applied to both BCI fields. The problem investigation is divided into three parts. The first analyses performance benchmarking of accuracies. A gap was found in how researchers reported their cross-validation methods, resulting in different BCI accuracy expectations. A thorough literature survey addresses the factors that contribute to these accuracy expectations. The second part examines the experiment setup factors that affect user experience when using BCIs, including workload and visual fatigue. Visual perception is explored to effectively design and implement a new BCI system that allows for user communication. The goal of this system is to ease the user fatigue caused by continued exposure to flickering lights, resulting in an improved user experience. The third explores the potential for emotion recognition to quantify violin quality, which could help conclude a long-lasting debate on the gold standard of violin quality. For this purpose, an emotion BCI system is examined to enable musical instrument classification. In the process, insights into how machine learning can be used to analyse instrument quality perception are given. This thesis includes literature surveys, critical analysis, questionnaire conducting, algorithm designs, and hardware development, with the aim of making BCIs a better field.

1.2 Thesis organisation

1.2.1 Introduction

The thesis begins with Chapter 2 which provides the necessarily background on the EEG and signal processing methods that are relevant to the thesis. This is followed in Chapter 3 by investigating cross-validation methods and their impact on performance inflation. In addition, a comparison between classifiers is conducted to test for the most suitable for the scope of thesis. Moreover, Chapter 4 discusses
the use of visual perception to enable communication. The aim of this new paradigm is to reduce user fatigue associated with flickering lights in common paradigms and, therefore, improve user experience. Finally, Chapter 5 introduces a novel emotion recognition system capable of classifying violin types. The aim of this system is to provide a proof of concept that different quality violins could be classified using EEG.

1.2.2 Chapter 3: Performance robustness

The aim of this chapter is to enhance robustness of the field by explaining for the first time why performances in the literature significantly vary, and provide a clear route to overcome this. This helps future studies to become more reliable. This is achieved in two steps. Firstly, by investigating the effect of different cross-validation methods, which has not been done before in BCI analysis, as different studies have shown variations in accuracies between 60% and 90%, using the same data and algorithms. Therefore, a critical analysis on the literature is conducted in order to re-implement these studies and conclude the root cause of these variations. The impact of this analysis extends to all chapters of the thesis as it enables correct implementation of classification. Secondly, various classifiers are tested in order to establish the most suitable for the scope of this research. The algorithm, which utilises state-of-the-art methods is developed to classify datasets representing both BCI fields. This algorithm is proven applicable to the two BCI systems discussed in Chapters 4 and 5. It involves combining Riemannian geometry with a fast yet highly accurate logistic regression (LR) classifier.

1.2.3 Chapter 4: User experience as a measure of performance

In the literature, the vast majority of studies use classification accuracy as an indicator of BCI performance. However, for BCIs to reach the market, there is a need to include usability as a measure of performance. This will not only ensure positive experience, but will prevent user fatigue and hence improve classification accuracies. This chapter introduces a new method to improve BCI user experience by reducing the fatigue associated with flickering lights. An ERP-based BCI system that enables communication without the use of flickering lights is developed.
for this purpose. More specifically, the system comprises a visual BCI paradigm that utilises brain behaviour associated with face recognition to enable communication. A survey is used to validate the proposed value of convenience in this new paradigm. Furthermore, the new paradigm is tested using a commercial-grade EEG recorder. To accurately detect stimuli onsets, a new wireless synchronisation circuit is designed and implemented which solves a technical challenge associated with low-cost EEG.

1.2.4 Chapter 5: Quantifying violin quality

The debate on how to define musical instrument quality has persevered for a significant time period, and has not been conclusively settled. A number of researchers have attempted to investigate physical and acoustic characteristics of specific instruments, for example, violins. Others have associated legendary makers, such as, Stradivarius with good quality. In this chapter, an emotion BCI system is developed to classify violin types. This system, which relies on subconscious behaviour rather than opinion, can potentially act as an objective measure for musical instrument quality, contributing to the debate of how to define the gold standard of quality. The system uses a web-based format to maximise user experience. The musical clips vary in valence and arousal levels to ensure independence from emotions. The preliminary tests involve the classification of valence and arousal levels. The main tests involve classification of violin types that expands the current understanding of brain behaviour that allowing for quantification of quality. This chapter includes opinion surveys, experiment design, EEG tests and analysis, and brain behaviour investigation.

1.2.5 Summary

This thesis’s contributions include new algorithms, solutions, and concepts presented in Chapters 3-5. A critical literature review and several study replications were completed to properly benchmark different setups and algorithms. A new system was designed and implemented to reduce BCI-user fatigue by enabling BCI communication without flickering screens, which is validated by a survey. Lastly, a novel experiment to classify violin types using EEG was completed to contribute
in understanding brain behaviour related to quality perception. The key objectives of the thesis are summarised as follows:

1. Investigate, in a rigorous manner, the problem of stated accuracies in the classification algorithms and provide clear recommendation on the best practices.

2. Design and implement a new BCI paradigm based on face and non-face recognition using commercial-grade EEG to reduce user fatigue.

3. Perform novel investigations using emotion BCI into perceptions of violin types.

References


Chapter 2

An overview of visual and emotion brain-computer interfaces: EEG hardware and signal processing

2.1 Introduction

BCI systems are comprised of several steps in order to achieve a useful output (see Figure 2.1).

Firstly, the EEG recorders produce raw data, which require significant signal processing before producing commands and information. EEG recorders are produced at different quality levels associated with various costs, which necessitates trade-offs when making a choice. Secondly, the data need to be converted into a more recognisable form by extracting relevant features. Some features are useful for event-related potential (ERP)-based BCIs, such as time-points, and some are useful for emotion-based BCIs, such as frequency domain. Thirdly, features require spatial filtering in order to obtain the most useful information and reduce redundancy. Fourthly, the information is then passed to a machine-learning classifier that trains on the data and tests it several times in order to produce an average output.

Figure 2.1: BCI architecture involving various steps like recording the data using an EEG, followed by signal processing such as feature extraction, spatial filtering, and machine learning. The output of the system defines the application of the system.
accuracy. This accuracy is the key performance indicator for BCI systems. It has a great impact on how we interpret the results of the experiments, and therefore taking extra care to not produce bias in calculating these accuracies is essential. This point is directly related to the first novel contribution in this thesis of appropriately benchmarking accuracies for more robust testing, discussed in Chapter 3. Finally, the output of the system determines its applications. The output is basically a mental state, which could be used for communication, as in Chapter 4, or to help investigate brain response to a specific stimulus in order to understand its behaviour, as in Chapter 5. A typical BCI experiment setup is illustrated in Figure 2.2.

There are various BCI paradigms that exploit EEG to achieve useful applications. In this introduction, a few paradigms that are relevant to the thesis are discussed as background information. Active BCI systems such as ERP-based P300-speller, steady-state visually-evoked potentials (SSVEP), and motor imagery (MI), are used to control other systems or communicate, similar to the work discussed in Chapter 4. Passive BCIs, such as emotion recognition systems, investigate brain behaviour under specific stimulus or mental states, for a more insight-providing goal, similar to the work discussed in Chapter 5. Figure 2.3 provides schematics on examples of how these 4 paradigms operate. Firstly, The P300-speller paradigm is a type of ERP-based BCIs. ERPs are waveforms represented by a letter, P or N, followed by a number. P represents a positive potential whereas N represents a negative potential. The number could be a single digit representing the order of a peak or trough, or three digits representing the latency in milliseconds of the peak or trough [2]. The P300 ERP is a positive peak around 300 milliseconds after a stimulus onset related to decision making (see Figure 2.4). Other ERPs include N170, visible in the visual cortex and common in face recognition [3], and N400, which is visible near centro-parietal electrode locations and common in responses to words and other meaningful stimuli [4]. The P300-speller BCI system uses a screen of sequentially flashing characters to enable communication, and brain behaviour is elicited when the intended character is stared at. It works by flickering complete rows and columns of characters. When the subject is staring at the flashing row and column containing the intended character, it induces P300 in the visual cortex. Knowing the row and the column reduces down the options to the intended character. Secondly, a SSVEP-based BCI has a similar setup to
Figure 2.2: Typical BCI setup involving 32 electrode locations, a screen for stimuli, EEG recorder, and cap setup. The 32 electrode locations follow the 10-10 international electrode placement system. Abbreviations: F: frontal, P: parietal, C: central, T: temporal, O: occipital. Cap and EEG recorder pictures obtained from [1]
Figure 2.3: BCI paradigms. (a) P300-speller allows communication by staring at characters. (b) SSVEP allows decision making by starting at flickering figures. (c) MI allows communication by imagine hand movement. (d) Emotion recognition of music expressing happy or sad emotions

the P300-speller. It enables BCI communication by having a user stare at a screen with boxes flashing at different frequencies. A user makes a decision by staring at the corresponding box. This decision is recognisable because the flashing box’s frequency is detectable in EEG. P300-speller and SSVEP share the fact that staring at flickering lights is needed, which causes user fatigue. This user-fatigue problem is tackled in Chapter 4 by introducing a new paradigm based on non-flickering images. Thirdly, a MI-based BCI enables communication by having the user imagine moving certain body parts, for instance, right- and left-hands, although some studies have expanded this 2-class classification to 4-class by including tongue and foot movement [5], [6]. Imagined hand movement causes decreased EEG activity at 8-12 Hz and 18-26 Hz over the motor cortex (near electrode positions C3 and C4) in contralateral hemispheres of the brain [7]. This could be used to control a drone to turn left or right by using mental commands [8]. Unlike P300-spellers and SSVEPs, MI needs individual training to achieve motor imagery [9]. Finally,
Figure 2.4: Examples of event-related potentials (ERPs): P1 is the first peak after stimulus, N170 is a negative potential that occurs 170 milliseconds after stimulus, and P300 is a positive potential that occurs 300 milliseconds after the stimulus.

an emotion recognition system is often utilised to predict a user’s mood or emotive perception of stimuli. A common application is to predict valence and arousal levels associated with musical clips, which helps assure that an intended musical perception is delivered. Another application is to investigate consumer brain behaviour in marketing campaigns for quality optimisation. To identify emotional states, frequency-domain information tests 8-13 Hz frontal asymmetries in the brain. An application of emotion recognition is introduced in Chapter 5.

This chapter analyses the quality of several EEG recorders and summarises recent developments in BCI signal processing methods. Section 2.2 covers both research- and commercial-grade EEG recorders, which is important because this study has used both types. Sections 2.3-2.5 address advancements in signal processing methods, including feature extraction and spatial filtering, and provide detailed classification algorithms. The purpose of this literature review is to provide the background information necessary to understand all the methods and contributions made in this thesis. The mathematical models explained in this chapter were implemented in Python using the open-source tools MNE, Scikit-learn, Keras, and PyRiemann. MNE [10] is an EEG pre-processing open-source Python library that enables analysis and visualisation of raw data. The library includes models, such as independent component analysis (ICA), for spatial filtering. Scikit-learn [11] is a very common machine-learning toolbox on Python: its feature extraction algorithms include Welch’s PSD and covariance matrices; its feature selection and spatial filtering tools include principle component analysis (PCA) and common spatial patterns (CSP); its classification algorithms include logistic regression (LR) and support vector machine (SVM); and its other validation methods include cross-
validation, grid-search for parameter tuning, and accuracy metrics such as the area under the curve (AUC). Keras [12] is a neural network implementation library. PyRiemann [13] is a Riemannian-geometry-based feature extraction, spatial filtering, and classification library. It enables xDAWN ERP-oriented spatial filtering, Riemannian CSP, and Riemannian classification. One of the contributions of this thesis is using these functions in novel and more accurate ways rather than redeveloping existing libraries.

2.2 EEG recorders

2.2.1 Introduction

This section provides an overview of different quality EEG recorders. There are different factors that affect the usability of an EEG recorder. Firstly, the quality of electrodes and their impact of the signal-to-noise ratio. This requires consideration of the appropriate recorder. For example, the rise of affordable commercial-grade EEG recorders has enabled many researchers to study BCI applications widely. However, this setup has raised valid concerns in the performance of different BCI systems. In commercial-grade EEG recorders, electrode and conductive solution quality impacts the signal-to-noise ratio and therefore results in lower classification accuracies as will be seen in this section. Secondly, the number of channels (electrodes) affects the accuracy and, therefore, requires the right use. Thirdly, commercial-grade recorders have provided convenience by enabling a quick setup with the use of less sophisticated consumables such as saline rather than a conductive gel, which comes with a price in signal quality. Finally, different EEG recorders come with different costs, and this deeply affects their availability. This section discusses the advantages and disadvantages of each EEG recorder type. The objective of this section is to familiarise the reader with both research- and commercial-grade EEG recorders because of their relevance to the other chapters.
2.2.2 Research-grade active electrode EEG recorders: Brain Products ActiChAmp and Biosemi Active Two

Brain Products’s ActiChAmp is a research-grade EEG recorder with active electrodes known for their low noise [14], [15]. Active electrodes amplify and digitise signals locally and then transmit data to the recording unit in the EEG. This reduces noise and outperforms passive electrodes. Centrally amplified EEGs read from passive electrodes and then transmit non-amplified analogue signals through cables, which are prone to motion artefacts and interference [16]. As a result, the transmitted noise is also amplified and digitised. The ActiChAmp uses up to 160 active electrodes with choices in multiples of 32 and connects with up to 8 auxiliary peripherals to enable other bio-signal sensors, such as galvanic skin response and respiration. The recorded EEG signal has a 24-bit resolution at a sampling rate of 100 KHz. However, this high sampling frequency is not always needed. According to the Nyquist theorem, the lowest sampling rate to preserve information is double the frequency band to be read [17]. In most BCI applications, frequencies above 60 Hz are not needed. Thus, a mere sampling rate of 120 Hz is sufficient. This EEG recorder is not popular because of its high cost. A simple Google Scholar database search of this EEG name returned only 440 studies that mentioned using it. One of the main downsides of this EEG recorder is that it takes a long time, on average 35 minutes to set up an experiment and, on average 15 minutes to clean afterwards [18]. This is because the electrodes require skin preparation to ensure low-impedance. Long experiments also cause user- and investigator-fatigue. The ActiChAmp was included because it is used in Chapter 5 of this thesis.

On the other hand, Biosemi’s Active Two is the most popular research-grade EEG recorder. A simple Google Scholar database search of this EEG name returned more than 5100 studies mentioning the Active Two. It features an active electrode system with up to 280 channels and can record EEG, electrocardiography (ECG), and electromyography (EMG). It also features a sampling rate of 16 KHz with a 24-bit resolution. While it features similar qualities as the ActiChAmp, there were no studies found comparing these different research-grade EEGs. The Active Two features in Chapter 3 of this thesis.
2.2.3 Commercial-grade EEG recorder: Emotiv EPOC+

Emotiv’s EPOC+ is the most common EEG recorder. A Google Scholar database search of this EEG name returned more than 5600 mentions for it. It is common for several reasons. Firstly, it is affordable, costing less than 5% of either research-grade EEG recorder mentioned in the previous section. Thus, the EPOC+ is a starting point for many researchers [19]. Secondly, it is fast, taking less than 5 minutes to setup [20], unlike research-grade recorders. Thirdly, its size makes it easy to carry, unlike research-grade recorders. Finally, it has 14 recording electrodes, more than the majority of other commercial-grade EEGs. In total, it has 14 fixed-location electrodes – AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4 – plus 2 reference electrodes at P3 and P4. Although these electrode locations, unlike cap-based electrodes, cannot be changed, they cover all sides of the brain with less resolution. The EPOC+ features a 2048 Hz internal sampling frequency, which is down-sampled to 256 Hz, with a 14-bit resolution. There are two downsides to this EEG recorder. Firstly, its poor-quality contact pads reduce the signal-to-noise ratio [19]. The saline used to wet these pads evaporates quickly [19], which degrades the data quality gradually and is therefore unsuitable for long experiments. Secondly, it features no hardware triggering channel, which causes difficulties with ERP-based stimuli that require accurate knowledge on stimulus onsets. However, this problem is tackled by designing a synchronisation solution, which will be discussed in Chapter 4. The EPOC+ is recommended for quick EEG experiments, such as SSVEP, P300-spellers, and passive BCIs. The EPOC+ is utilised in Chapter 4 of this study.

2.2.4 Summary

This section covered three EEG recorders, all of which were relevant in this study. Firstly, the research-grade ActiChAmp, used in Chapter 5, which features high-quality active electrodes and increases the signal-to-noise ratio. However, it requires a long set-up. Secondly, the research-grade Active Two, further discussed in Chapter 3, that remains the most common research-grade EEG. Thirdly, the commercial-grade EPOC+, further discussed in Chapter 4, that is easy and fast to set-up, although its electrode contact material has been criticised for causing a
Another contribution this thesis makes to the BCI field, is in comparing the performances of different EEG recorders, with respect to quality and price points, and introducing a new testing method so that classification results can be more robustly compared between different studies with different recorders. The effect of variations in setup and algorithm on performance is discussed in detail in Chapter 3.

2.3 Feature extraction

2.3.1 Introduction

Although there are several ways of extracting features, the most commonly used features are time-points and frequency-band features [21]. This section outlines feature extraction methods relevant to this thesis. The algorithm developed in Chapter 3 uses features discussed in this section, such as power spectral density (PSD) and covariance matrices. Other feature extraction methods which are repeatedly mentioned in the literature review in Chapter 3, such as time-points, statistical features, are also included in this section.

2.3.2 Time-points

Time points are raw EEG data from all electrodes that have been filtered or pre-processed to be fed directly to a classifier. Depending on their application, these features typically use a band-pass filter to remove data of no relevance. They can also be down-sampled to lower frequencies, saving memory and increasing computational speed. Notably, reducing the number of features increases the efficiency of classifiers and necessitates less parameter optimisation [21]. Time-points features are useful for ERP-based classification, in which class information lies in the time-domain and is represented by temporal variations. This information is typically time-locked, meaning that it appears an exact number of milliseconds after a stimulus onset as was discussed in Section 2.1. Time points are not specifically utilised in this thesis but they are included for completeness as they are the simplest form.
of features and are still widely used in BCI as will be seen in Chapter 3.

However, there are disadvantages to using time-points features. They often contain a large number of values, some of which are redundant and lack the relevant information to differentiate between classes [21]. Furthermore, a large number of features results in a large number of parameters that need tuning [21], creating extra computational complexity. Moreover, a large number of features increases the chance of over-fitting, especially when the number of training trials is small [21]. More features also result in more information to process, meaning that detecting the relevant features can be difficult. These problems can be addressed with the use of different features, such as statistical features, discussed in Section 2.3.5, or by spatial filtering, discussed in Section 2.4.

2.3.3 Power spectral density (PSD)

PSD information is typically extracted by a Fourier transformation and represents the power or energy of a particular frequency or oscillations in the brain. This information is typically expressed in frequency bands, known as delta (0-4 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz), and gamma (30-100 Hz). PSDs are widely used in BCIs to extract band oscillatory activity, represented by changes to signal rhythms and amplitudes. Band-power features are considered the gold standard in passive BCIs [21]. Notably, frequency-domain information is related to MI, SSVEP, and emotion BCIs and are used in Chapter 3 in extraction of emotion features from EEG.

The first step of calculating PSD is a Fourier transformation, achieved using the discrete Fourier transform (DFT). The real and imaginary parts, as defined in [17], are calculated by

$$X_R[k] = \frac{2}{N} \sum_{i=0}^{N-1} X_i \cos(2\pi ki/N) \quad (2.1)$$

$$X_I[k] = \frac{2}{N} \sum_{i=0}^{N-1} X_i \sin(2\pi ki/N), \quad (2.2)$$

where $X_i, i = 0, 1, ..., N - 1$ is the input signal, $N$ is the number of samples, and $k$ is the relevant harmonic to be calculated. Once the real and the imaginary coefficients are obtained, the spectrum displayed is the magnitude of both vectors using
the identity

\[ X[k] = \sqrt{X_R[k]^2 + X_I[k]^2}. \]  

(2.3)

The computational complexity of the DFT is \( n \times k \), which is high [17]. The fast Fourier transform (FFT) can have an impact because this may take a significant time to compute for long signals. The concept behind it is that signal is decomposed into even and odd sequences, forming a number \( M \) of 2-point recordings. Next, calculate the DFTs of these \( M \) 2-point sets and recompose the \( M/2 \) 4-point sets from the \( M \) 2-point sets. This process is repeated until the full-length DFT is acquired. The power spectral density is then computed by scaling the squared absolute value of the signal’s FFT to get the power for each frequency component. Thus, this makes the output

\[ P_x = \frac{1}{F_s \times N} |\text{FFT}|^2, \]  

(2.4)

where \( P_x \) is the power spectral density, \( F_s \) is the sampling frequency, and \( N \) is the number of samples [22]. PSDs can be calculated in different ways, such as with Welch’s method [23]. The data in this method is split into smaller epochs with a certain level of overlapping. Each segment’s FFT is then calculated and squared. The PSD values are the average values of all segments. Other methods to extract PSDs, such as the spectrogram and the Wigner-Ville distribution methods, could be found in this comparative study [24].

### 2.3.4 Covariance matrices

Covariance matrices are another method of feature extraction, one related to the method of common spatial patterns (CSPs), discussed later in Section 2.4.3. A covariance matrix represents the energy relation between all channels for the duration of the data trial. It assumes that energy spatial distribution is fixed during a mental state. The covariance matrix’s diagonal parts represent the variance of electrode energy and the off-diagonal parts represent the covariance between pairs of electrodes. Such information has been successful when the classes are differentiated by brain asymmetries, such as right- and left-hand movement imagery and emotion alpha frontal asymmetries. Covariance matrices are used Chapters 3, 4, and 5.
The sample covariance matrix (SCM) of an input signal $\mathbf{X}$ of length $N$ is defined by [25] as

$$SCM = \frac{1}{N} \mathbf{X}\mathbf{X}^T.$$  \hfill (2.5)

Another type of ERP features are extracted using covariance matrices. However, since ERP-based time-domain features have smaller changes in amplitude compared to the background EEG, covariance matrices of a single trial cannot directly extract them [26]. Therefore to use covariance matrices in time-domain analysis, a method such as that used in [26] can be implemented: by calculating the average of each class and then finding the covariance between these averages and the trial data. This enables the extraction of temporal information by comparing the energy of a single trial to the energy of the average ERPs. The covariance matrix will indicate the similarity between the trial data and both classes. ERP extraction can be maximised by applying spatial filters to increase the signal-to-noise ratio, which will be discussed in Section 2.4. The following step, described in [27], extracts the SCM of each filtered trial $\mathbf{Z}_l$ by concatenating it with the average $\hat{\mathbf{P}}_k$ of each class $k$:

$$\hat{\mathbf{Z}}_l = \begin{bmatrix} \hat{\mathbf{P}}_0 \\ \hat{\mathbf{P}}_1 \\ \mathbf{Z}_l \end{bmatrix}.$$  \hfill (2.6)

Next, the SCM feature matrix of each trial is calculated as

$$\mathbf{Z}_{SCM} = \frac{1}{N - 1} \hat{\mathbf{Z}}_l \hat{\mathbf{Z}}_l^T,$$  \hfill (2.7)

where $N$ is the number of samples in a single trial. This method is robust to noise, and outliers, such as data covariance, are calculated with an original, known ERP average [27]. This method has shown to be competent, as it won a classification competition of ERP-based BCI data [28].

### 2.3.5 Statistical features

Statistical features are often used to reduce the number of time-point features. They are also used to extract energy information associated with different pairs of electrodes. Furthermore, information theory-based features, such as entropy, are widely used in addition to higher order crossings (HOCs). An example of the
importance of these feature was seen in an emotion recognition survey [29] which stated that 23.8% of emotion classification studies have used statistical features, second only to DFT features. Statistical features are included because of their relevance to literature survey in Chapter 3.

The most common features in this survey will now be described. The arithmetic mean, as defined in [30], [31], of a signal $X_i$, $i = 1, 2, ... N$ with size $N$ samples from each channel is calculated by

$$\mu = \frac{1}{N} \sum_{i=1}^{N} X_i. \quad (2.8)$$

The variance, as defined in [30], of a signal $X_i, i = 1, 2, ... N$ with size $N$ samples represents how far the data are spread from the mean $\mu$. This is calculated by

$$\sigma = \frac{1}{N} \sum_{(i=1)}^{N} (X_i - \mu)^2. \quad (2.9)$$

The energy, as defined in [30], of a signal $X_i, i = 1, 2, ... N$ with size $N$ samples is calculated by

$$E = \sum_{i=1}^{N} X_i^2. \quad (2.10)$$

The mean energy, as defined in [30], of a signal $X_i, i = 1, 2, ... N$ with size $N$ samples is calculated by

$$ME = \frac{1}{N} \sum_{i=1}^{N} X_i^2. \quad (2.11)$$

Sample entropy, as defined in [30], represents the randomness of a signal $X_i, i = 1, 2, ... N$ with size $N$ samples and is calculated by

$$En = \sum_{i=1}^{N} X_i^2 \log_2 (X_i^2). \quad (2.12)$$

Hjorth parameters have been widely used [31], and these include activity, mobility, and complexity, as defined in [31]. The activity of a signal $X_i, i = 1, 2, ... N$ with size $N$ samples is a measure of sample energy and is calculated by

$$Act = \frac{1}{N - 1} \sum_{i=1}^{N} (X_i - \sigma(X_i))^2. \quad (2.13)$$
The mobility of a signal $X$ is defined as the ratio of its slope to its amplitude, calculated by

$$Mob = \sqrt{\frac{Act(XX')}{Act(X)}},$$

(2.14)

where $X'$ is the difference operator for discrete signals. $XX'$ is the equivalent of the differential operator $\frac{dx}{dt}$ of a continuous signal $x$. The complexity of a signal $X$ is the change in its frequencies and indicates the similarity between a signal and a pure sine wave, calculated by

$$Comp = \frac{Mob(XX')}{Mob(X)}.$$  

(2.15)

Another feature uses higher order crossings (HOCs), which represent the number of zero-crossings in a signal [32]. A zero-crossing is the point in time where the amplitude value crosses zero, calculated by

$$D_i = \sum_{i=2}^{N} (Y_i - Y_{i-1})^2,$$ 

(2.16)

where

$$Y_i = \begin{cases} 
1 & X_i \geq 0 \\
0 & X_i < 0. 
\end{cases}$$ 

(2.17)

Therefore, HOCs are defined as

$$HOC = [D_1, D_2, ..., D_M].$$ 

(2.18)

HOCs have shown advancements over other feature extraction methods [33], [34].

These features are sometimes taken from smaller windows known as epochs (for example see [31], [35]). This helps account for variations in time, where energy bands can increase or decrease per epoch, and therefore provides more temporal accuracy. There are other statistical features of importance that have been used in BCIs. For example, the peak-to-peak mean is the arithmetic mean of the vertical length from the top value of a cycle to the top value of the following cycle [35].
2.3.6 Digital wavelet transformation (DWT)

Another feature utilised in frequency-domain information extraction, is DWT. However, DWTs are not used in this research but are included for completeness as they are widely used, which will become apparent in Chapter 3.

The DWT of a continuous signal $X(t)$, as defined in [36], is

$$DWT(i, k) = \frac{1}{\sqrt{|n^j|}} \int_{-\infty}^{\infty} X(t) \psi \left( \frac{t - 2^j k}{n^j} \right) dt,$$

(2.19)

where $\psi$ is the wavelet function, $2^j$ is a scaling parameter, and $2^j k$ is the time localisation parameter. A common wavelet function is Daubechies levels 4 and 5 [36], [37]. DWT is useful for extracting frequency information whilst preserving time information [37], an advantage over the normal DFT.

2.3.7 Summary

Different feature extraction methods were discussed. Time-points data, associated with ERP-based BCIs, usage is commonly used. However, it can include redundant data that increase complexity and confuse classifiers. This could be solved using methods like spatial filtering. PSD is extracted by calculating the DFT and then dividing the data into smaller windows using Welch’s method. Covariance matrices indicate the energy variance in each electrode and the energy relation between pairs of electrodes. They could also be used to extract ERP information by combining information about different classes and the data trial. Statistical features – such as mean, variance, energy, and entropy – can reduce dimensionality. Finally, DWTs can be used to extract frequency-domain information with higher temporal resolution.

2.4 Spatial Filtering

2.4.1 Introduction

This section focuses on spatial filtering methods. Spatial filtering involves increasing the signal-to-noise ratio by locating areas of interest in the brain and max-
iming information extraction with respect to signals from other areas. Spatial filtering has several analytical benefits. Firstly, it reduces the number of channels and features that are fed to the classifier, which decreases error and computation time. Secondly, it removes noise from areas of interest by reconstructing the data into virtual channels formed from linear combinations of real channels. Finally, it operates with either knowing the areas of interest, as in independent component analysis (ICA), or knowing the class of each trial, as in common spatial patterns (CSPs). In the former case, unsupervised ICA can remove all data from other areas and improve classification, while supervised CSP can help identify areas of interest and filter testing data based on information from training data. Supervised spatial filtering means that the class of each trial is known and used to create the filters. In contrast, unsupervised spatial filtering means that the classes are not used to create the filters in a training step. In this section, brief descriptions of two conventional, wide-spread methods, unsupervised ICA and supervised CSP, are given and then updates on the state-of-the-art xDAWN and modified CSPs, including filter-bank- and Riemannian-geometry-based CSPs, are presented. For both time- and frequency-domain information, the use of spatial filtering before feature extraction in BCIs is very common [21].

2.4.2 Unsupervised ICA

ICA was used in Chapters 3 and 4 and was included here as background information. To obtain the ICA equation, as described in [38], assume n recorded signals $X_n$ that are linear mixtures of n independent components $S_n$:

$$X_j = a_j1s_1 + a_j2s_2 + \ldots + a_jns_n, j = 1, \ldots, n$$ (2.20)

or

$$X = As,$$ (2.21)

where $s$ is a vector containing components $s_1, \ldots, s_n$ and $A$ is a vector containing components $a_{ij}$. The objective is to solve for $A$ knowing components $s_i$ are independent and must have a non-Gaussian distribution. Solving this problem, known
as whitening, can be done using the covariance matrix’s decomposition:

\[ VDV^T = E[\hat{X}\hat{X}^T], \tag{2.22} \]

where \( V \) is the matrix of the orthogonal eigenvectors and \( D \) is a diagonal matrix with the corresponding eigenvalues. Whitening is done by multiplying the recorded signal with the transformation matrix \( P \):

\[ \hat{X} = PX, \tag{2.23} \]

where

\[ P = VD^{1/2}V^T. \tag{2.24} \]

Therefore, one way of computing ICA is by maximising the non-gaussianity of the whitened mixture of signals. A quantitative measure of non-gaussianity of the signal \( \hat{X} \) using the kurtosis method, defined by [39] as

\[ \text{kurt}(\hat{X}) = E[\hat{X}^4] - 3(E[\hat{X}^2])^2 \tag{2.25} \]

Non-gaussianity is defined by the absolute value of the kurtosis, where a gaussian variable is has a value of zero and non-gaussian variable has a value greater than zero.

Now to maximise the absolute value of kurtosis, in practice we need to begin with a random vector \( w \). Then compute the direction of the value of the kurtosis of \( w^T\hat{X} \) is growing strongly in and move the vector \( w \) in that direction. This method is known as the gradient algorithm. As a result, the gradient of the absolute value of kurtosis of \( w^T\hat{X} \) is defined as

\[ \frac{\partial|\text{kurt}(w^TX)|}{\partial w} = 4 \text{sign(kurt}(w^T\hat{X})(E[\hat{X}(w^T\hat{X})^3] - 3w||w||^2) \tag{2.26} \]

However, this method is slow, and therefore ‘FastICA’ or fast fixed-point algorithm using kurtosis computes a more efficient fixed-point iteration. Equating the
gradient of kurtosis defined in 2.26 with \( \mathbf{w} \) results in

\[
\mathbf{w} \propto (E[\hat{\mathbf{X}}(\mathbf{w}^T \hat{\mathbf{X}})^3] - 3||\mathbf{w}||^2 \mathbf{w})
\]

(2.27)

This equation suggests a fixed-point algorithm giving \( \mathbf{w} \) a new value of

\[
\mathbf{w} \leftarrow E[\hat{\mathbf{X}}(\mathbf{w}^T \hat{\mathbf{X}})^3] - 3\mathbf{w}
\]

(2.28)

After every fixed-point iteration, \( \mathbf{w} \) is divided by its norm to remain on the constraint set i.e. \( ||\mathbf{w}|| = 1 \). The final value of \( \mathbf{w} \) gives the independent components. Therefore the de-mixing equation of the ICA becomes the simple of projection of the de-mixing matrix \( \mathbf{w} \) on the original signal \( \hat{\mathbf{X}} \) resulting in a matrix \( \hat{\mathbf{S}} \) of estimated sources with the same number of channels

\[
\hat{\mathbf{S}} = \mathbf{w}^T \hat{\mathbf{X}}.
\]

(2.29)

This method converges quickly and reliably. In addition, it requires to parameter tuning or learning rate adjustment, which makes it easy to use.

ICA is very common in early analyses. However, it is mostly visually applied to individual experiments and, consequently, takes a long time to complete. While it can remove all non-relevant signals and increase the signal-to-noise ratio, it requires information about the areas of interest and the specific kind of stimuli used. However, it can also remove noise associated with eye- and muscle-movements due to their distinctive features. ICA was used in this study for this purpose and is further discussed in Chapter 4.

### 2.4.3 Supervised CSP

CSP is included because it used in Chapters 4 and 5. CSP is known to maximise the variance of two-class signal matrices. A class in this context refers to the type of stimulus that differentiates it from other stimuli. For example, high and low valence are two classes of emotion and pictures of a face and a non-face are also two classes. The process, explained in [40] is essentially to project the matrices using a transformation process into a smaller space with the maximum eigenvectors of
one class in addition to the minimum of the other class and vice versa. The first stage of applying CSP is to find the covariance for the 2 classes. The variance $R$ of a single trial $X$ for each class is given by

$$R_X = \sum \frac{XX^T}{tr(XX^T)}, \quad (2.30)$$

where the input $X$ represents all data for a single trial, taken in by all channels. The columns represent samples, and the rows represent channels. The composite spatial covariance for 2 classes is given by

$$R = R_1 + R_{c2}, \quad (2.31)$$

where $R_1$ and $R_2$ are the average normalised variances for the 2 classes. $R$ can be factorised into

$$R = U_0 \Lambda U_0^T, \quad (2.32)$$

where $U_0$ is the matrix of eigenvectors and $\Lambda$ is the diagonal matrix of eigenvalues. Next, define the whitening matrix as

$$P = \sqrt{\Lambda} U_0^T \quad (2.33)$$

and use it to transform the average covariance matrices into

$$SCM_1 = PR_1 P^T, \quad (2.34)$$

$$SCM_2 = PR_2 P^T. \quad (2.35)$$

As a result, $SCM_1$ and $SCM_2$ share common eigenvectors and the sum of the corresponding eigenvalues for the matrices will be the identity matrix $I$. Thus,

$$SCM_1 = U \Lambda_1 U^T, \quad (2.36)$$

$$SCM_2 = U \Lambda_2 U^T, \quad (2.37)$$
\[ \Lambda_1 + \Lambda_2 = I. \]  

The eigenvectors with the largest eigenvalues for \( \text{SCM}_1 \) have the smallest eigenvalues for \( \text{SCM}_2 \) and vice versa. To separate the variances of these classes, the largest eigenvalues for both classes are needed. Then, a transformation of the original data occurs using

\[ W = U^T P, \]  

where \( W \) is the projection matrix and can be used as

\[ X_{\text{new}} = WX. \]

CSP methods are common, especially in MI-based BCIs, largely due to their robustness in maximising the class difference related to different brain areas, such as in right- and left-hand imagery movement [21]. An example of spatially filtered EEG is shown in Figure 2.5.

2.4.4 Supervised filter-bank- and Riemannian-geometry-based CSPs

A couple of methods based on CSPs are discussed in this section. Firstly, the filter-bank common spatial patterns (FBCSP) method was first introduced by [41]. It exploited the use of filter bank and feature selection to improve CSP performance.
The method is illustrated in Figure 2.6. The filter bank comprises several band-pass filters that isolate the features of each frequency band. CSP projection is then calculated for each frequency band separately, which is followed by mutual information-based feature selection. A total of 9 band-pass filters cover the range of 4-40 Hz, and the concatenation of these CSP-filtered bands is then fed to a feature-selection algorithm to choose the most suitable feature. This method is common due to its performance in passive BCIs [21]. Other methods based on this were developed (for example see [42]).

![Figure 2.6](image)

Figure 2.6: Overall process of the filter-bank common spatial patterns (FBCSP) involving applying CSPs to various frequency bands then choosing the most suitable using feature selection. Figure obtained from [41]

Secondly, an advanced CSP method utilising Riemannian geometry, described by [28], [43], [44], is discussed. Let $X_l$ be a input data of a trial $l$. A SCM is defined as

$$ SCM(X) = \frac{1}{T-1} X_l X_l^T. $$

(2.41)

To apply Riemannian CSP spatial filtering, the mean of a class’s covariances is needed. Given symmetric positive definite (SPD) matrices $P_1, P_2, ..., P_L$ the Riemannian mean $RM$, as defined in [44], is

$$ RM(P_1, P_2, ..., P_L) = \arg \min_{P \in P(N)} \sum_{i=1}^{I} \delta_R^2(P, P_i), $$

(2.42)

where $\delta_R(P_1, P_2)$ is the Riemannian distance between two symmetric positive def-
inite SPD matrices, $P_1$ and $P_2$ in $P(n)$, and is given by

$$\delta_R(P_1, P_2) = \left[ \sum_{l=1}^{L} \log^2 \frac{\lambda_l}{1 - \lambda_l} \right]^{1/2},$$

(2.43)

where $\lambda_l$ are the eigenvalues. Riemannian distance approximates Euclidean distance in the Euclidean feature space, which measures the separability of classes in this feature space. Next, the discriminative power of each spatial filter $w_j$ is related to its associated eigenvalue by the relationship

$$(A_1 + A_2)^{-1} A_1 w_j = \lambda_j w_j,$$

(2.44)

where $A_1$ and $A_2$ are the Riemannian mean matrices of classes 1 and 2, respectively. Thus, the filter is chosen based on what values maximise the quantity $|\lambda_j - 0.5|$. The filtered output $\hat{X}$ therefore is obtained by the direct projection of SCM on each trial with the spatial filter $W$ that maximises the covariance between the classes:

$$\hat{X} = W^T SCM(X).$$

(2.45)

This method has shown improvements in classification accuracy over standard CSPs [44], and is used in Chapters 3, and 5.

### 2.4.5 Supervised xDAWN

This method, as described by [27], [45], calculates the average $P_k$ of the training trials $X_l, l = 1, 2, ..., L$, where $L$ is the total number of training trials for each class $k$ separately:

$$P_k = \frac{1}{L} \sum_{l=1}^{L} X_l.$$  

(2.46)

Thus, for each class $k$, the spatial filter $w$ which maximises the time-domain features of the class average $P_k$ over all training data trials $X$ follows a generalised Rayleigh quotient equation:

$$\hat{w} = \arg_{w} \max \left( \frac{w^T P_k P_k^T w}{w^T X X^T w} \right).$$  

(2.47)
Therefore, the m eigenvectors that correspond to the largest eigenvalues in the decomposition of \( P_k P_k^T (XX^T)^{-1} \) are taken from each class to form \( w_k \). As a result, the spatially filtered trial \( Z_l \) is constructed by the linear projection of the spatial filter \( W = [w_0, w_1] \) with the original trial data \( X_l \):

\[
Z_l = W^T X_l.
\]

To illustrate the effect on filtered data, Figure 2.7 shows two diagrams of ERPs spatially filtered to increase signal to noise ratio on the visual cortex.

![Figure 2.7: xDAWN spatial filtering. (a) Shows the projection of first component of the filter \( W \) plotted on user’s scalp increasing signal to noise ratio of ERPs in visual cortex in the back of the brain (b) shows the second component of the filter with noise that could be removed. Figure obtained from [45] with modifications](image)

xDAWN good performance is due to its ability to maximise the separation between classes by increasing the signal-to-noise ratio and removing redundant data [21]. xDAWN is also used in this research and is discussed in Chapters 3 and 4.

### 2.4.6 Summary

The conventional, wide-spread spatial filtering methods ICA and CSP were explained. ICA is an unsupervised spatial filtering method. For a signal comprising several channels that form linear combinations of different sources, it separates those sources by solving the eigenvalue decomposition of these channels’ covariance matrices, attenuating noise. CSP is a supervised spatial filter that required labels for each trial to identify its class for training. It uses the eigenvalue decomposition of the class-average covariance matrices to construct virtual channels that maximise the signal-to-noise ratio for one class and minimise it for the other. The filters are constructed from the training data and applied with linear multiplication to the testing data.
In addition, state-of-the-art methods were also covered. These are used in this thesis. xDAWN maximises information extraction from ERP-based data by calculating the class-averages from training data and then constructing each class-average covariance matrix to follow the same process in CSP. Filter-bank CSP separates the data into frequency-bands before applying CSP and then uses mutual information for feature selection. Finally, Riemannian-geometry-based CSP filters the data by maximising the Riemannian distances of the class-average covariance matrices. These methods were all proven to improve classification accuracy.

2.5 Classification

2.5.1 Introduction

In this section, various classification algorithms are discussed because of their relevance to various other parts of the thesis. These include short descriptions of conventional, wide-spread methods such as Logistic regression (LR) and Support vector machines (SVM), which have either been used in other chapters or are mentioned frequently. In addition, longer explanations of advanced methods, such as deep neural networks (DNNs) and Riemannian-geometry-based classifiers are included because they have been reported to have the best performance in state-of-the-art studies. In addition, various configurations of cross-validations which affect classification accuracy reporting are described in addition to how these accuracies are calculated and tested for statistical significance.

2.5.2 Logistic regression (LR)

LR is a simple linear classifier that utilises statistical regression analysis to give a class probability if certain features or conditions are provided. Let $\hat{X}_l$ be the feature vector of trial $l$, where $y_l$ is the label 0 or 1 indicating its class. The hypothesis or probability of the occurrence function $h$, as defined in [46], [47], is a sigmoid with an output of range $[0, 1]$, calculated by

$$h_\theta(\hat{X}_l) = \frac{1}{1 + \exp(-\hat{X}_l \times \theta^T)}.
$$ (2.49)
where $\theta$ is a weight vector of same length as the feature vector $\hat{X}_l$. The output, meaning the class label, is defined as

$$y_l = \begin{cases} 
0 & h_\theta(\hat{X}_l) < 0.5 \\
1 & h_\theta(\hat{X}_l) \geq 0.5.
\end{cases} \quad (2.50)$$

The function is trained by minimising the cost function with the parameters of the weight vector $\theta$. LR is used in Chapters 3, 4, and 5.

### 2.5.3 Support vector machines (SVM)

SVM identifies the optimal hyperplane for classification using supervised learning as seen in Figure 2.8.

![Figure 2.8: Representation of hyperplane optimisation separating two classes (green and red). The optimal hyperplane separates the classes with as close to equal distances as possible.](image)

The equation, as described in [48], of the hyperplane is given by

$$< w, \phi(X_l) > + b = 0, \quad (2.51)$$

where $< w, \phi(X_l) >$ represents the dot product of $w$ and $\phi(X_l)$. For training trial input data $X_l, l = 1, 2, ..., L$ and class label $y_l$, there are 2 SVM methods to solve this problem: Primal and Dual SVM. Using the Primal method for lower dimensions [49], the values for $w$ and $b$ must be found that minimises $||w||$ for all data points $(X_l, y_l)$ such that

$$y_l(< w, \phi(X_l) > + b) \geq 1. \quad (2.52)$$
When \( y_i (< w, \phi(X_i) > + b) = 1 \), \( \phi(X_i) \) is called the support vector. Thus, a new trial \( X_z \) is classified using

\[
\text{class}(X_z) = \text{sign}(< w, \phi(X_z) > + b),
\]

(2.53)

where the sign (positive or negative) of the term \( (< w, \phi(X_z) > + b) \) indicates whether it is the first class or the second class.

Now, using the Dual SVM method for higher dimensions [49], \( \alpha \) must be minimised for

\[
\frac{1}{2} \alpha^T Q \alpha - e^T \alpha
\]

(2.54)

subject to \( y^T \alpha = 0 \), where \( 0 \leq \alpha_l \leq C, l = 1, 2, ..., L \), \( C \) is a regularisation parameter, \( e = [1, ..., 1]^T \) is the vector of all ones, \( Q \) is an \( L \) by \( L \) semidefinite matrix \( Q_{lj} = y_l y_j K(X_l, X_j) \) and \( K(X_l, X_j) = \phi(X_l)^T \phi(X_j) \) is the kernel function. As a result, the optimal hyperplane equation \( w \) becomes

\[
w = \sum_{l=1}^{L} y_l \alpha_l \phi(X_l)
\]

(2.55)

Thus, a new trial \( X_z \) is classified using

\[
\text{sign}(w^T \phi(X) + b) = \text{sign}\left(\sum_{l=1}^{L} y_l \alpha_l K(X_l, X) + b\right)
\]

(2.56)

where the sign (positive or negative) of the term indicates whether it is the first class or the second class. The variables \( y_l, \alpha_l \) and \( b \) are trained on data then stored for prediction use later on.

SVM is a very popular BCI classification algorithm due to its performance in complicated BCI systems [21], [50]. SVM is discussed in Chapter 3.

2.5.4 Neural networks (NNs)

A NN is a machine learning method that utilises mathematical neurons to create classifiers and learn features directly from the data [51]. Deep learning uses multiple layers other than only input and output layers to learn these features (e.g.
A traditional simple feedforward neural network is the MLP, which comprises an input layer, hidden layers, and an output layer [53]. The neurons at the outputs are connected to the inputs of the following layer inputs. The final layer output is a single neuron to determine class of the input. The structure of the neural network is important and therefore requires careful selection [50]. However, the number of neurons in each layer is best determined by trial and error [53]. An excessive number of neurons or insufficient may cause problems of overfitting [53]. At hidden layers, each neuron $j$ equals the sum of the input signal samples $X_i$ multiplied by weighting scalars $w_{ji}$. The output of each neuron, as defined by [53], is described by

$$y_j = f\left(\sum w_{ji}X_i\right)$$

(2.57)

where $f$ is an activation function utilising the weighted summations of the inputs, which could be a simple threshold function [53]. The training process involves min-
imising the value of the error $E$ and adjusting weighting scalars $w_{ji}$

$$E = \frac{1}{2} \sum_j (y_{dj} - y_j)^2$$  \hspace{1cm} (2.58)

Where $y_{dj}$ is the desired value of output neuron $j$ and $y_j$ is the actual output value of that neuron during training.

The other common type of DNNs is CNNs. CNNs are neural networks that consist of at least one convolutional layer [21]. The information in CNNs flows into the input layer, the hidden layers, and then the output layer. The features and the classifier are extracted from the data in these layers [21]. CNNs learn local data patterns by passing information into their convolutional layers [54], and a stacked group of these layers forms a classifier that can be trained and then used to predict unknown classes. Each convolutional layer is followed by non-linear or pooling layers that aggregate the information into a small number of outputs, such as in Max pooling [55]. To expand their learning, CNNs need large amounts of training data. They initially learn local low-level features and increasingly gain enough complexity to learn global features [54]. Notably, CNNs can learn non-linear features by using their convolutional processes and non-linearities to gain complexity [54].

In the past ten years, NNs have outperformed other state-of-the-art BCI classification methods. Examples of studies exploiting NNs include the following [55]–[57]. However, there are drawbacks to using NNs. Firstly, they need extensive parameter tuning and optimisation [21]. Secondly, they need many trials to train, which is usually difficult to obtain in BCI applications. Finally, they are computationally complex and require long periods of time to both train and predict, which affects their ability to perform in real-time. Although NNs can demonstrably classify complicated sets of data, problems exist with parameter justification and reported accuracy benchmarking in the literature. Most BCI studies did not clearly justify the suitability of their chosen parameters [21]. In addition, some studies poorly explained their data-split and cross-validation methods, both of which significantly affect performance. These problems will be extensively discussed in Chapter 3.
2.5.5 Riemannian-geometry-based minimum distance to the mean classification

The Riemannian mean and Riemannian distances were described in Equations 2.42 and 2.43 in Section 2.4.4.

These tools can be exploited to classify symmetric positive definite (SPD) matrices directly in their manifolds. In Riemannian geometry, the manifold, representing the original feature space, is a smooth curved space that can be locally and linearly approximated, where this linear approximation at each point is called the tangent space. The tangent space at a Riemannian manifold is equipped with an inner product varying between different points and this results into a non-Euclidean distance between any two points. This distance could geometrically be calculated using the Riemannian distance, which is adapted to the geometry of the curved space, instead of the Euclidean distance, which is a extrinsic distance [21]. The Riemannian distances enables the use of a scalar product on each point to find the angles and lengths of curves on the manifold [58]. These terms are illustrated in Figure 2.10.

![Figure 2.10: Representation of the relationship between SPD manifold and tangent space at feature matrix G. The shortest path between two points \((C_1, C_2)\) is the Riemannian distance. Figure obtained from [28]](image)

A simple classification process uses the minimum distance to the mean (MDM), as defined in [25]. This process includes computing the Riemannian mean \(C^k\) of each class \(k\) from the training trial’s covariance matrices \(SCM_i\). Then, the class of each new trial is distinguished from the closest Riemannian mean of the testing trial covariance matrix \(SCM_j\):

\[
k = \arg \min_k \delta^2(SCM_j, C^k)
\]  

(2.59)
Riemannian-geometry-based methods work equally well on all BCI paradigms, as they allow for the extraction of both time- and frequency-domain features, with the only difference being how the matrices are mapped onto the tangent space. If both methods are combined, the paradigm and feature types do not have to be known in advance.

Moreover, Riemannian geometry allows for the classification of features in their matrix form, such as covariance matrices, instead of their vector form, which is useful in BCI design because it preserves the spatial information of adjacent-electrode data points. The Riemannian-geometry-based method is robust to noise, and its improved generalisation capability is conditioned with a relatively small number of electrodes. Notably, its computational complexity increases cubically with the number of electrodes used [21]. However, this can be avoided by spatial filtering, a very common method that reduces the number of channels and increases the signal-to-noise ratio [21], as was discussed in Section 2.4.

Riemannian geometry can also be used for spatial filtering and feature extraction before standard operations, such as SVM and LR for classification, are applied. By doing so, this both maximises information in feature extraction using Riemannian geometry and exploits complex decision functions using standard classifiers. This method is discussed in more detail in Chapter 3.

### 2.5.6 Cross-validation methods

Classification reportings can vary from one test to another. A typical method to test a classifier is to train it on a portion of the data, and test it on another portion by predicting classes of new data. Therefore, the number of correct predictions out of all predictions determines the system accuracy. However, to achieve a more reliable test, a cross-validation method could be used. In cross-validation, the testing process is repeated several times with different divisions of training and testing data. This allows the classifier to produce several prediction accuracies and report their average as the system accuracy. There are various forms of cross-validation and data split methods that are used to test classifiers. Firstly, the use of subject-dependant and subject-independent data splits. Subject-dependant takes the data of one participant for training and tests it using various data split meth-
methods as discussed below. Subject-independent takes the data from all participants except one for training and then tests on this one participant, before repeating this process for all participants. This method is also known as leave-one-subject-out (LOSO) cross-validation. Secondly, cross-validation could be implemented in two ways that have an impact on the system accuracy: nested and non-nested testing. The essence of the difference between those two methods is whether the classifier reports accuracy on new data or not. An explanation of this problem in non-BCI contexts is provided by [59].

![Diagram of Nested and Non-nested Cross-Validation](image)

**Figure 2.11:** Nested and non-nested cross-validation methods. The diagram shows nested cross-validation tests on new data while non-nested tests on validation data. Both processes are repeated n time to calculate the average accuracy.

With respect to Figure 2.11, to understand the difference between nested and non-nested cross-validation methods, imagine splitting the data (X) into 2 parts, training (X_{train}) and testing (X_{test}), where the former is what the classifier is trained on and the latter is what the classifier is tested on. In nested cross-validation, the training data (X_{train}) is split into 2 smaller parts to improve the parameters; one is for tuning the classifier parameters, and the other is for validating the param-
eters. This process is repeated several times. The parameters with the highest accuracy are then used to examine the system performance on the testing part ($X_{test}$). On the other hand, in non-nested cross-validation, the parameters are tuned without splitting the data ($X_{train}$) into 2 smaller parts, and without testing the new data, the best validation accuracy is then reported as the system accuracy.

The effect of these two methods was examined by [60]. The test involved reporting accuracies on the 1936 Iris flower dataset [61]. SVM classification was done with with several parameter optimisations. Figure 2.11 shows that the difference between the two methods varies between 0 and 2.5%. The impact of cross-validation methods on emotion recognition classification is discussed in Chapter 3.

**2.5.7 Area under the receiver operating characteristic curve (AUC)**

The AUC is an accuracy measure often used to represent classification performance [21]. It is measured by the total area below the receiver operating characteristic (ROC) curve. An AUC of 1 is flawless, and a natural random prediction of two classes is around 0.5. The ROC curve represents probability rates of false positive classification and true positive classification [62] (see Figure 2.13).
On their own, true positive accuracies lack the ability to represent performance when the number of trails for both classes is imbalanced. For example, if there were hundred trials with ten trials of class 0 and ninety trials of class 1, and the system predicts all hundred classes to be 1, then the system true positive rate is 90%, which clearly is not representative of performance [21]. ROC analysis is beneficial as it provides a single accuracy number that represents the classification performance [64].

2.5.8 Statistical significance using the t-test

Statistical significance is determined by the p-value, which could be calculated using the t-test. The t-test, explained by [65], assumes a normal distribution for a sample of data. However, the sample variance is needed to calculate t distribution. The variance of a sample $x$ is defined as

$$\sigma = \frac{1}{N} \sum_{(i=1)}^{N} (x_i - \mu)^2. \quad (2.60)$$

where $\mu$ is the arithmetic mean, and $N$ is the length. Therefore, the t value of a sample from another sample, which could for example be classification sample from chance sample, is defined as

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \quad (2.61)$$
Figure 2.14: Graphs of p-values obtained from value of t and df, where df = n-1 as shown in the figure. Figure obtained from [66]

To confirm a statistical significance between two samples, or between a sample and natural chance values, the two samples have be independent. Now the degree of freedom (df) of a sample is defined as $N - 1$. Thus, to find the p-value, the values of t and df has follow the graphs shown in Figure 2.14, or from tabulated values of the same graphs [67]. As a result, a p-value less than 0.05 means that an improbable event has occurred [66], which indicates its statistical significance.

2.5.9 Summary

This section described several well-known classification methods: LR, and SVM. Firstly, LR uses a weight function to limit the output of the hypothesis function to the range between 0 and 1. Secondly, SVM defines a separation kernel, known as hyperplane, based on the support vectors.

More advanced methods were also described, including the neural network classification method CNN, which trains a number of layers on data using at least one convolutional layer and other non-linearities. While CNNs have been successful, they require many training trials. A method based on Riemannian geometry was also explained. It classifies trials using the shortest distance from the average training trials. Notably, Riemannian-geometry-based methods have outperformed many other state-of-the-art methods. This gives the mathematical background which will be used in Chapters 3, 4, and 5.

Moreover, nested and non-nested cross-validation methods were discussed. These methods affect the accuracy reporting of classification. Nested cross-validation
tests the classification on new data and therefore has lower accuracy but a more reliable and generalisable performance. Non-nested cross-validation searches for the best parameters and report accuracy on the same data without testing new data, which causes higher accuracies but less generalisation. Classification accuracies are often reported in AUC, which provides information on both true positive and false positive classifications. Finally, accuracies are tested for statistical significance using the t-test, which compares a sample of accuracies against the whole population.

2.6 Chapter Summary

This chapter provided the required background to understand the contributions made in the rest of the thesis. Section 2.1 summarised four common BCI paradigms: P300-speller, SSVEP, MI, and emotion recognition.

Section 2.2 discussed various EEG recorders. Research-grade EEG recorders, such as Brain Products ActiChAmp and Biosemi Active Two, take longer to setup and therefore affect convenience. However, they provide the best data quality, which significantly improves BCI performance. Commercial-grade EEG recorders, such as Emotiv EPOC+, take a shorter amount of time to setup and are user friendly. This comes with the disadvantage of poorer data quality.

Section 2.3 described BCI feature extraction methods. Time-point features are used after being filtered and down-sampled. To reduce the number of features involved, spatial filtering is often applied. For the same reason, statistical features – such as mean, variance, sample entropy, and HOCs – can be used. PSD features are obtained form FFTs of the signal, which split the data into smaller windows, before computing. Covariance matrices also enable feature extraction, as calculating the average SCM of the trials provides information on the energy variance of each electrode, and the energy relation between each pair of electrodes. Such information is useful when brain energy asymmetries exist between classes. DWTs extract frequency-domain information similar to DFTs but with higher temporal resolution.

Section 2.4 described spatial filtering methods that increase the signal-to-noise
ratio and reduce the number of features, which helps avoid classifier confusion caused by a large number of features. Conventional methods include unsupervised ICA and supervised CSP. ICA removes all non-relevant information from areas other than the area of interest. CSP maximises the difference between two classes based on energy distributions within the brain. The eigenvalue decomposition of the covariance matrices forms virtual channels that maximise the signal-to-noise ratio of each class. Other state-of-the-art spatial filtering methods included xDAWN and modified CSPs. xDAWN is similar to CSP except that it decomposes the covariance matrices of the class-average ERP grand trial instead of the class-average covariance matrix. xDAWN is particularly useful for ERP-based BCIs as it accounts for time variant information. Filter-bank CSP divides the data into smaller frequency bands and then computes their CSPs. Mutual information feature selection is then used to choose the most-suitable frequency bands for classification. Furthermore, Riemannian-geometry-based CSPs utilise the Riemannian mean instead of the arithmetic mean and Riemannian distance instead of eigenvalues to identify spatial filters. Notably, filter-bank and Riemannian-geometry-based CSPs outperform conventional CSPs.

Section 2.5 described both conventional and state-of-the-art classification methods. The LR decision function determines the class of a trial based on a weight function with the same length as the feature vector. SVM separates classes with a hyperplane that divides the area between the support vectors in half. The function that describes the hyperplane is the kernel. In addition, advanced DNNs and Riemannian-geometry-based methods were discussed. CNNs are an application of DNNs that use convolution layers in addition to non-linearities to train their networks and predict new data. While CNNs can have an outstanding performance in non-BCI work, this is not necessarily the case in BCI work because of their need for many training trials. The Riemannian-geometry-based classification method uses Riemannian distances to choose the closest mean to the covariance matrix of a trial, and it can compete with other state-of-the-art methods. Different cross-validation methods were also discussed. Nested cross-validation results in lower but more generalisable accuracies, while non-nested cross-validation results in higher accuracies but they are less generalisable. Accuracies are typically reported in AUC and tested for statistical significance using the t-test.
Now the different type of methods have been introduced, it is possible to evaluate the factors which affect their performance in Chapter 3.

References


Chapter 3

An evaluation of cross-validation and classifier methods for visual and emotion BCIs:

Recommendations for robust performance

3.1 Introduction

The aim of this chapter is to increase data analysis robustness for BCI development in order to help researchers benchmark their work with the literature. This not only in itself a contribution, but it also helps evaluating the work discussed in Chapters 4 and 5. It was achieved in two steps. Firstly, Section 3.2 analyses classification accuracies in emotion recognition by conducting a systematic review of studies utilising neural networks (NNs) and investigating the cause of variations in accuracies in the range 60-90%. Subsequently, re-implementing NN classification in nested and non-nested cross-validation methods to prove that many studies report inflated performances with non-nested cross-validation methods. Secondly, once an appropriate benchmarking is established with nested cross-validation, Section 3.3 discusses algorithm development based on state-of-the-art methods and tested with logistic regression (LR), support vector machine (SVM), and NN classifiers. This algorithm testing is done in various configurations including two datasets
representing ERP and emotion EEG, and different channel combinations to estab-
lish the best performing classifier.

3.2 Critical analysis of cross-validation methods and their impact on NN performance inflation

3.2.1 Introduction

Accuracy of a classification algorithm is critical in determining the performance and reliability of a system. This creates motivation for researchers to boost accuracies. In specific, the use of non-nested cross-validation method, which was discussed in Section 2.5.6, could be used to inflate the accuracy of a classifier [1]. An example is NNs that are believed to be structured to train on testing data and therefore inflate performance [2].

In this section, emotion recognition algorithms are investigated to explain variations that are found between studies on classification accuracy testing their algorithms on the same dataset. For example, in the year 2018, some studies have claimed accuracies of 86% [3], 92% [4], and 95% [5], while others utilising similar classifiers have claimed accuracies of 60% [6], 62% [7], and 62% [8]. This may result from the implementation of non-nested cross-validation methods that allow for testing information to be leaked to classifiers during training, as was defined in Section 2.5.6. In this section, variations between BCI studies will be examined, validated, and quantified in the hope of encouraging researchers to accurately report their cross-validation methods. Such findings could facilitate a route to a more robust understanding of BCI accuracies and help manage expectations for new readers in the field. Section 3.2.2 describes the motivation behind the analysis. Section 3.2.3 describes the methods for literature survey and classification. Sections 3.2.4 and 3.2.5 provide the results and discussion, respectively.

3.2.2 Motivation

Several studies in the literature have provided detailed methods for their algorithms, and these clearly indicate leak of information in cross-validation. Informa-
tion leakage can invalidate the predictive capabilities of the model by including information that would not typically be available in the training data. For example, a study [9] that used SVMs and declared a 2-class classification accuracy of 73% also calculated a 63% chance level for predicting two classes. This is obtained by randomly allocating class-labels to trials 0 and 1, running the algorithm multiple times, and then calculating the average accuracy. A chance level of 63%, rather than 50%, suggests that some of the information could have been leaked to the classifier. One possible cause for this is the use non-nested cross-validation methods. Another study [10] employing NNs reported an accuracy of 75% and provided a pseudo code of its training and testing process. This code clearly showed the use of test data as validation data to optimise parameters followed by the testing of the same validation data (see Figure 3.1). These non-nested cross-validation methods resulted in an unfounded high accuracy.

![Algorithm 1 Pseudocode for Convolutional Neural Model](image)

Figure 3.1: A study utilising a non-nested cross-validation method. Algorithm 1, Lines 21, 22 in [10] indicate the use of validation data as testing data.

Although the aforementioned findings do prove that the use of non-nested cross-validation is a potential explanation for these variations in accuracy, it is essential to eliminate all other possible reasons. To enable this, an investigation into emotion recognition studies was done utilising a literature survey accomplished by [11]. The literature survey covered all emotion recognition studies in the period
2009-2016. It analysed studies for aspects like equipment setups, such as EEG hardware and signal processing methods such as feature extraction, classification, and filtering. The data obtained from the survey helped provide essential information. Firstly, it allowed grouping the studies by type of EEG recorder. This is important to determine whether research-grade and commercial-grade data quality is the explanation for these variations. Secondly, the cross-validation data split is either subject-dependent or -independent. This also significantly affects accuracy. Thirdly, the type of classifier could be examined to evaluate whether it is the reason behind the aforementioned variations. Table 3.1 summarised the information obtained from [11]. There were 23 studies reporting the use of Biosemi Active Two research-grade EEG recorder and 10 studies reporting that of Emotiv commercial-grade EEG recorder. The average accuracy of the research-grade studies was 72% with 14 studies claiming accuracies above 70%. On the other hand, the average accuracy of the commercial-grade EEG was 76% with 7 studies claiming accuracies above 70%. These findings suggest that there is no relation between the type of EEG recorder and the reported accuracy. In fact, as opposed to what is expected from a commercial-grade EEG, the reported accuracy is higher than research-grade EEG. The variations of research- and commercial-grade are discussed in Chapter 4. It is also worth mentioning that these studies vary in data-split methods, such as subject-dependent and subject-independent methods, which also eliminates this as a reason. Moreover, for the commercial-grade EEGs, 8 out of 10 studies reported using the same SVM classifier. In contrast, for the research-grade EEGs, 10 out of 23 studies reported using SVMs, with 4 out of these 10 having accuracies between 70.0% and 95%. These results eliminate the possibility that the classification algorithm was the sole cause of improved performance. This evidence motivated us to quantify the effect of both methods. A significant difference between the two methods indicates that many of these studies reporting high accuracies may be utilising non-nested cross-validation and therefore confounding the rest of scientific community.

3.2.3 Methods

In this section, a systematic literature review is introduced to find studies with a common testing dataset to examine various accuracies associated with feature ex-
Table 3.1: Snapshot of emotion recognition studies utilising two different EEG recorders. Data obtained from literature survey [11]. Reference numbers from the original study.

<table>
<thead>
<tr>
<th>Reference</th>
<th>EEG Recorder</th>
<th>Data Split</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>RF</td>
<td>50</td>
</tr>
<tr>
<td>71</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>SVM</td>
<td>52</td>
</tr>
<tr>
<td>55</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>LDA</td>
<td>57</td>
</tr>
<tr>
<td>49</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>SVM</td>
<td>59</td>
</tr>
<tr>
<td>111</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>LS-SVM</td>
<td>61</td>
</tr>
<tr>
<td>70</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>SVM</td>
<td>62</td>
</tr>
<tr>
<td>29</td>
<td>Biosemi Active Two</td>
<td>Dep</td>
<td>SVM</td>
<td>63</td>
</tr>
<tr>
<td>78</td>
<td>Biosemi Active Two</td>
<td>Dep</td>
<td>SVM</td>
<td>63</td>
</tr>
<tr>
<td>22</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>C4.5</td>
<td>68</td>
</tr>
<tr>
<td>62</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>kNN</td>
<td>70</td>
</tr>
<tr>
<td>83</td>
<td>Biosemi Active Two</td>
<td>Dep</td>
<td>NB</td>
<td>71</td>
</tr>
<tr>
<td>75</td>
<td>Biosemi Active Two</td>
<td>Dep</td>
<td>LR</td>
<td>71</td>
</tr>
<tr>
<td>84</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>kNN</td>
<td>72</td>
</tr>
<tr>
<td>105</td>
<td>Biosemi Active Two</td>
<td>Dep</td>
<td>SVM</td>
<td>73</td>
</tr>
<tr>
<td>40</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>QDA</td>
<td>77</td>
</tr>
<tr>
<td>45</td>
<td>Biosemi Active Two</td>
<td>Dep</td>
<td>BLDA</td>
<td>80</td>
</tr>
<tr>
<td>116</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>kNN</td>
<td>84</td>
</tr>
<tr>
<td>7</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>SVM</td>
<td>85</td>
</tr>
<tr>
<td>72</td>
<td>Biosemi Active Two</td>
<td>Dep</td>
<td>kNN</td>
<td>96</td>
</tr>
<tr>
<td>101</td>
<td>Biosemi Active Two</td>
<td>Ind</td>
<td>ML-SVM</td>
<td>96</td>
</tr>
<tr>
<td>57</td>
<td>Biosemi Active Two</td>
<td>Dep</td>
<td>LDA</td>
<td>100</td>
</tr>
<tr>
<td>113</td>
<td>Emotiv</td>
<td>Ind</td>
<td>SVM</td>
<td>36</td>
</tr>
<tr>
<td>115</td>
<td>Emotiv</td>
<td>Dep</td>
<td>SVM</td>
<td>58</td>
</tr>
<tr>
<td>96</td>
<td>Emotiv</td>
<td>Dep</td>
<td>SVM</td>
<td>62</td>
</tr>
<tr>
<td>65</td>
<td>Emotiv</td>
<td>Dep</td>
<td>SVM</td>
<td>77</td>
</tr>
<tr>
<td>95</td>
<td>Emotiv</td>
<td>Both</td>
<td>SVM</td>
<td>84</td>
</tr>
<tr>
<td>32</td>
<td>Emotiv</td>
<td>Ind</td>
<td>SVM</td>
<td>85</td>
</tr>
<tr>
<td>103</td>
<td>Emotiv</td>
<td>Dep</td>
<td>SVM</td>
<td>86</td>
</tr>
<tr>
<td>64</td>
<td>Emotiv</td>
<td>Ind</td>
<td>kNN</td>
<td>87</td>
</tr>
<tr>
<td>68</td>
<td>Emotiv</td>
<td>Dep</td>
<td>AdaBoost</td>
<td>93</td>
</tr>
<tr>
<td>120</td>
<td>Emotiv</td>
<td>Ind</td>
<td>SVM</td>
<td>94</td>
</tr>
</tbody>
</table>

Acronyms: Data split: subject-dependant (Dep), subject-independent (Ind). Classifiers: Bayesian linear discriminant analysis (BLDA), Kth nearest neighbour (KNN), linear discriminant analysis (LDA), logistic regression (LR), least-square support vector machines (LS-SVM), multi-class support vector machines (ML-SVM), Naive Bayes (NB), quadratic discriminant analysis (QDA), random forest (RF), SVM support vector machines (SVM).

Based on this literature review, a combination of feature extraction and classifier is re-implemented with both nested and non-nested cross-validation methods, with the aim of quantifying the effect of these configurations.

**Literature Review**

To determine the best possible accuracy that could be obtained with both cross-validation methods and calculate the chance level for each method, a new literature survey was completed of studies testing their algorithms on the same dataset.
The literature survey included all studies since 2017 that both cited the music-based dataset DEAP [12] and tested their algorithms on this dataset. A total of 221 studies were found, within which 59 studies reported testing their algorithms on the DEAP dataset. Each study was examined for 4 pieces of information. Firstly, the type of data split that was used in cross-validation. This is primarily subject-dependent or subject-independent. Accuracies with subject-independent are expected to be lower than subject-dependant. Secondly, the feature-extraction method that was utilised in the algorithm. Thirdly, the type of classifier that was used. Finally, the maximum reported average accuracy. Some studies performed more than one test, and therefore only their personal highest accuracy was selected.

The dataset used for this analysis was obtained from [12]. It included the data of 32 subjects watching 40 one-minute music video clips that were categorised into different levels of arousal and valence. The 40 clips were short-listed from 120 clips using a survey that aimed to choose the most suitable clips. The experiments involved an EEG recorder with 32 electrodes sampling at a frequency 512 Hz. In addition, 12 other peripheral channels were used to record physiological signals. Each subject rated the clips in terms of arousal, valence, and dominance on a continuous 1-9 scale. The definitions of high and low levels were extracted from the participants’ personal ratings. The pre-processed data (down-sampled to 128 Hz) were publicly available to download along with the participants’ ratings.

Re-implementation

To test the hypothesis that the use of non-nested cross-validation is what causes the large gap in reported accuracies, an algorithm was developed to test both types of cross-validation. The algorithm utilised PSD feature extraction and a NN for classification. PSD was chosen because of its popularity and NN was chosen because of its inflated reported accuracies, as will become apparent in the results. The algorithm consists of several steps to produce a final system accuracy. See Algorithm 1. Background information on these steps was provided in Chapter 2. Lines 2 and 3 are concerned with loading the data into the environment. The features were obtained by calculating the PSD of the EEG data, as seen from Line 11. The multi-layer perceptron (MLP) NN-classifier employed in this study had a sequential model presented by Lines 4-10. MLP was chosen because of its small
Algorithm 1  Nested vs non-nested cross-validation methods

1: procedure GetAccuracy
2: \$X \leftarrow \text{data (trials, channels, samples)}\$
3: \$y \leftarrow \text{labels}\$
4: \$\text{model}_\text{MLP} \leftarrow \text{Sequential()}\$
5: \$\text{model}_\text{MLP}.\text{add(Dense(1.0*\text{n_feature}),activation='relu')}\$
6: \$\text{model}_\text{MLP}.\text{add(Dense(0.5*\text{n_feature}),activation='relu')}\$
7: \$\text{model}_\text{MLP}.\text{add(Dense(0.25*\text{n_feature}),activation='relu')}\$
8: \$\ldots\$
9: \$\text{model}_\text{MLP}.\text{add(Dense(1), activation='sigmoid')}\$
10: \$\text{model}_\text{MLP}.\text{compile(loss='binary_crossentropy', optimizer='rmsprop', metrics='auc')}\$
11: \$\text{model} \leftarrow \text{make_pipeline(PSD, model}_\text{MLP)}\$
12: \$X_{\text{train}}, X_{\text{test}} \leftarrow \text{data_split}(X)\$
13: \$\text{model}_\text{non_nested}.\text{fit}(X_{\text{train}}, \text{validation data}=X_{\text{test}}, y)\$
14: \$\text{non_nested score} \leftarrow \text{model}_\text{non_nested}.\text{cross_val_score}(X_{\text{test}}, y)\$
15: \$\text{model}_\text{nested}.\text{fit}(X_{\text{train}}, y)\$
16: \$\text{nested score} \leftarrow \text{model}_\text{nested}.\text{cross_val_score}(X_{\text{test}}, y)\$

the number of parameters that require optimisation. In general, classifiers with fewer parameters are expected to perform better for small number of trials [2]. Line 5 shows that the first layer had an input dimension equal to the number of features in a single trial, with the same number of output neurons as number of features with rectified linear units (RELU) activation. RELU activation was used to introduce non-linearity in the model, which makes it more practical for this application [10], [13]. The second layer had an output number of neurons equal to half of those in the previous layer, which would allow for enough information to be passed in from the previous layer. The third layer followed a similar structure, and so on. The final layer in Line 9 had an output shape of one neuron to produce a 1 or 0 for either class. Line 10 is responsible for compiling the model with the loss calculated using a binary cross-entropy metric, because of having two classes. The accuracy is calculated by the AUC, which was explained in Section 2.5.7. Line 12 splits the data into training and testing portions. The difference between nested and non-nested cross-validations lies in Lines 13-16. Non-nested cross-validation takes the variable \$X_{\text{test}}\$ as validation data, which means the model is going to train by knowing the testing data, which is where the leak of information occurs. In contrast, nested cross-validation has no access to testing data. The function \text{cross_val_score}() returns the accuracy for the corresponding model. The chance levels were calculated by feeding random class labels (0 or 1) to the classifier for training and testing instead of the true labels that represent high and low valence levels. The process then repeats the predictions 100 times. Chance levels are then computed as the average of accuracies of all 100 computations with random labels. The p-
values were calculated using the t-test, which was explained in Section 2.5.8. This analysis enabled feeding the true chance level to the t-test instead of hypothetical chance level of 50%.

### 3.2.4 Results

The results are divided into 2 parts. The literature survey and the re-implementation of the classification algorithm.

#### Literature Review

This section summarises the results of the systematic literature review study. The results include data splits, feature extraction, classifier, and accuracy for each paper included.

Table 3.2: Literature survey of studies testing their algorithms on the DEAP dataset.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Data split</th>
<th>Feature Extraction</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[14]</td>
<td>Dep</td>
<td>PSD</td>
<td>cTL</td>
<td>64</td>
</tr>
<tr>
<td>[15]</td>
<td>Dep</td>
<td>Raw</td>
<td>DNN</td>
<td>88</td>
</tr>
<tr>
<td>[9]</td>
<td>Dep</td>
<td>PSD</td>
<td>SVM</td>
<td>73</td>
</tr>
<tr>
<td>[16]</td>
<td>Dep</td>
<td>BPC,HOC</td>
<td>SVM</td>
<td>67</td>
</tr>
<tr>
<td>[17]</td>
<td>Dep</td>
<td>PSD,DE</td>
<td>LR,SVM</td>
<td>70</td>
</tr>
<tr>
<td>[18]</td>
<td>Dep</td>
<td>DWT</td>
<td>SVM</td>
<td>62</td>
</tr>
<tr>
<td>[10]</td>
<td>Dep</td>
<td>Raw</td>
<td>DNN</td>
<td>75</td>
</tr>
<tr>
<td>[19]</td>
<td>Dep</td>
<td>PCC</td>
<td>SVM</td>
<td>85</td>
</tr>
<tr>
<td>[20]</td>
<td>Dep</td>
<td>IMF</td>
<td>SVM</td>
<td>70</td>
</tr>
<tr>
<td>[5]</td>
<td>Dep</td>
<td>SE</td>
<td>KNN</td>
<td>95</td>
</tr>
<tr>
<td>[21]</td>
<td>Dep</td>
<td>DWT</td>
<td>KNN</td>
<td>84</td>
</tr>
<tr>
<td>[3]</td>
<td>Dep</td>
<td>IMF</td>
<td>SVM</td>
<td>86</td>
</tr>
<tr>
<td>[22]</td>
<td>Ind</td>
<td>T-RFE</td>
<td>SVM</td>
<td>75</td>
</tr>
<tr>
<td>[23]</td>
<td>Dep</td>
<td>LSTM</td>
<td>RNN</td>
<td>85</td>
</tr>
<tr>
<td>[24]</td>
<td>Ind</td>
<td>T-RFE</td>
<td>SVM</td>
<td>78</td>
</tr>
<tr>
<td>[6]</td>
<td>Ind</td>
<td>PSD,Statistical</td>
<td>SVM</td>
<td>60</td>
</tr>
<tr>
<td>[25]</td>
<td>Dep</td>
<td>MEMD</td>
<td>ANN</td>
<td>73</td>
</tr>
<tr>
<td>[26]</td>
<td>Dep</td>
<td>DWT</td>
<td>KNN</td>
<td>89</td>
</tr>
<tr>
<td>[27]</td>
<td>Dep</td>
<td>Raw</td>
<td>AdaBoost</td>
<td>90</td>
</tr>
<tr>
<td>[28]</td>
<td>Dep</td>
<td>PSD,Statistical</td>
<td>RBFNN</td>
<td>64</td>
</tr>
<tr>
<td>[29]</td>
<td>Dep</td>
<td>DWT</td>
<td>SVM</td>
<td>45</td>
</tr>
<tr>
<td>[30]</td>
<td>Dep</td>
<td>PSD</td>
<td>DNN</td>
<td>82</td>
</tr>
<tr>
<td>[31]</td>
<td>Dep</td>
<td>Raw</td>
<td>SVM</td>
<td>80</td>
</tr>
<tr>
<td>[32]</td>
<td>Dep</td>
<td>LSM</td>
<td>LSM</td>
<td>84</td>
</tr>
<tr>
<td>[33]</td>
<td>Dep</td>
<td>PSD,PSO</td>
<td>SVM</td>
<td>86</td>
</tr>
<tr>
<td>[34]</td>
<td>Dep</td>
<td>Raw</td>
<td>GBDT</td>
<td>75</td>
</tr>
<tr>
<td>[35]</td>
<td>Dep</td>
<td>Statistical</td>
<td>CNN</td>
<td>87</td>
</tr>
<tr>
<td>[36]</td>
<td>Dep</td>
<td>CNN</td>
<td>CNN</td>
<td>73</td>
</tr>
<tr>
<td>[37]</td>
<td>Dep</td>
<td>DWT</td>
<td>SVM</td>
<td>68</td>
</tr>
<tr>
<td>[38]</td>
<td>Dep</td>
<td>DWT</td>
<td>SVM</td>
<td>92</td>
</tr>
<tr>
<td>[39]</td>
<td>Dep</td>
<td>DWT</td>
<td>KNN</td>
<td>77</td>
</tr>
<tr>
<td>[40]</td>
<td>Dep</td>
<td>PSD, Statistical</td>
<td>SVM</td>
<td>73</td>
</tr>
</tbody>
</table>
Table 3.2 shows that only 12 (out of 59) studies reported accuracies below 70%. This indicates that there still variations amongst the same dataset. However, the vast majority of emotion recognition algorithms in the literature are above 70%. Regarding classifiers, 25 studies used SVMs, the average accuracy of those studies is 73%. Furthermore, 14 studies used NNs, with an average accuracy of 81%. NNs appear to have an inflated performance, which supports the concern of inflated performance raised by [2]. Regarding feature extraction, 17 studies used PSD features, while 15 used raw data, and 11 used DWTs. This indicates that PSD are the most used features. These two facts have motivated the re-implementation of NNs with PSD features, the results of which are discussed in the following section. In conclusion, the accuracies range from 62% to 99%. This suggests that there

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[7]</td>
<td>Dep</td>
<td>Statistical, HOC, FD</td>
<td>SVM</td>
<td>77</td>
</tr>
<tr>
<td>[7]</td>
<td>Dep</td>
<td>Raw</td>
<td>DNN,CNN</td>
<td>62</td>
</tr>
<tr>
<td>[41]</td>
<td>Dep</td>
<td>DWT</td>
<td>SVM</td>
<td>64</td>
</tr>
<tr>
<td>[8]</td>
<td>Dep</td>
<td>Raw</td>
<td>LSTM</td>
<td>62</td>
</tr>
<tr>
<td>[42]</td>
<td>Dep</td>
<td>FCC</td>
<td>KNN</td>
<td>90</td>
</tr>
<tr>
<td>[43]</td>
<td>Dep</td>
<td>PSD</td>
<td>SVM</td>
<td>63</td>
</tr>
<tr>
<td>[44]</td>
<td>Dep</td>
<td>PSD</td>
<td>SVM</td>
<td>82</td>
</tr>
<tr>
<td>[45]</td>
<td>Dep</td>
<td>MFI</td>
<td>CNN</td>
<td>75</td>
</tr>
<tr>
<td>[46]</td>
<td>Dep</td>
<td>Raw</td>
<td>CNN</td>
<td>83</td>
</tr>
<tr>
<td>[47]</td>
<td>Dep</td>
<td>PSD</td>
<td>DNN</td>
<td>87</td>
</tr>
<tr>
<td>[48]</td>
<td>Dep</td>
<td>Raw</td>
<td>CNN</td>
<td>78</td>
</tr>
<tr>
<td>[49]</td>
<td>Dep</td>
<td>PSD</td>
<td>LSTM</td>
<td>77</td>
</tr>
<tr>
<td>[50]</td>
<td>Dep</td>
<td>PSD,DWT</td>
<td>SVM</td>
<td>75</td>
</tr>
<tr>
<td>[4]</td>
<td>Dep</td>
<td>Raw</td>
<td>DTL</td>
<td>92</td>
</tr>
<tr>
<td>[51]</td>
<td>Dep</td>
<td>PSD</td>
<td>DNN</td>
<td>79</td>
</tr>
<tr>
<td>[52]</td>
<td>Dep</td>
<td>PSD,DWT</td>
<td>DNN</td>
<td>92</td>
</tr>
<tr>
<td>[53]</td>
<td>Dep</td>
<td>Raw</td>
<td>SVM</td>
<td>76</td>
</tr>
<tr>
<td>[54]</td>
<td>Dep</td>
<td>Raw</td>
<td>KNN</td>
<td>75</td>
</tr>
<tr>
<td>[55]</td>
<td>Dep</td>
<td>DWT</td>
<td>KNN</td>
<td>99</td>
</tr>
<tr>
<td>[56]</td>
<td>Dep</td>
<td>Raw</td>
<td>DNN</td>
<td>95</td>
</tr>
<tr>
<td>[57]</td>
<td>Dep</td>
<td>PSD</td>
<td>MESAE</td>
<td>76</td>
</tr>
<tr>
<td>[58]</td>
<td>Dep</td>
<td>Raw</td>
<td>DBN</td>
<td>78</td>
</tr>
<tr>
<td>[59]</td>
<td>Dep</td>
<td>SE</td>
<td>SVM</td>
<td>72</td>
</tr>
<tr>
<td>[60]</td>
<td>Dep</td>
<td>MEMD</td>
<td>SVM</td>
<td>71</td>
</tr>
<tr>
<td>[61]</td>
<td>Dep</td>
<td>PSD,Statistical</td>
<td>SVM</td>
<td>78</td>
</tr>
<tr>
<td>[62]</td>
<td>Dep</td>
<td>Raw</td>
<td>MNN</td>
<td>88</td>
</tr>
<tr>
<td>[63]</td>
<td>Dep</td>
<td>RE</td>
<td>RF</td>
<td>79</td>
</tr>
</tbody>
</table>

Acronyms: Data split: subject-dependant (Dep), subject-independent (Ind). Features: transfer recursive feature elimination (T-RFE), digital wavelet transform (DWT), fractal dimension (FD), higher order crossings (HOC), intrinsic mode functions (IMF), liquid state machines (LSM), multivariate empirical mode decomposition (MEMD), multidimensional feature image (MFI), power spectral density (PSD), particle swarm optimization (PSO), Renyi entropy (RE), sample entropy (SE). Classifiers: convolutional neural network (CNN), conditional transfer learning (cTL), deep belief network (DBN), deep neural network (DNN), dempster–shafer theory (DST), gradient boosted decision trees (GBDT), Kth nearest neighbour (KNN), logistic regression (LR), multiple-fusion-layer based ensemble classifier of stacked autoencode (MESAE), radial basis function neural network (RBFNN), random forest (RF), support vector machines (SVM).
Table 3.3: Nested and non-nested cross-validated classification of high and low calence using MLPs and PSD features for the DEAP dataset [12]

| Participant | Nested | | Non-nested | |
|-------------|--------| |------------| |
|             | Accuracy (AUC) | Chance (AUC) | Accuracy (AUC) | Chance (AUC) |
| 1           | 47      | 43   | 90          | 58          |
| 2           | 62      | 59   | 76          | 67          |
| 3           | 49      | 47   | 78          | 72          |
| 4           | 58      | 55   | 85          | 63          |
| 5           | 58      | 55   | 96          | 64          |
| 6           | 73      | 74   | 98          | 72          |
| 7           | 67      | 70   | 92          | 68          |
| 8           | 50      | 47   | 84          | 67          |
| 9           | 58      | 44   | 80          | 72          |
| 10          | 58      | 43   | 88          | 67          |
| 11          | 66      | 55   | 94          | 67          |
| 12          | 53      | 44   | 86          | 67          |
| 13          | 67      | 47   | 91          | 73          |
| 14          | 66      | 42   | 87          | 72          |
| 15          | 74      | 43   | 98          | 59          |
| 16          | 64      | 60   | 80          | 63          |
| 17          | 42      | 47   | 97          | 70          |
| 18          | 60      | 63   | 96          | 69          |
| 19          | 70      | 51   | 95          | 65          |
| 20          | 58      | 50   | 94          | 68          |
| 21          | 48      | 43   | 59          | 67          |
| 22          | 48      | 43   | 76          | 73          |
| 23          | 68      | 67   | 85          | 72          |
| 24          | 42      | 45   | 97          | 70          |
| 25          | 52      | 44   | 96          | 67          |
| 26          | 63      | 63   | 97          | 60          |
| 27          | 76      | 75   | 95          | 60          |
| 28          | 59      | 59   | 97          | 69          |
| 29          | 55      | 51   | 98          | 64          |
| 30          | 64      | 67   | 91          | 63          |
| 31          | 72      | 64   | 96          | 65          |
| 32          | 68      | 43   | 95          | 67          |

Average 59.69 53.33 89.57 66.82
Standard deviation 9.29 10.14 8.87 4.29
P-value 0.0004 0.0001

could be a problem with how these accuracies are reported, especially in cross-validation methods.

Re-implementation

This section summarises the results obtained from re-implementing the algorithm comprising PSD features and NN classification using both nested and non-nested cross-validation methods. The accuracies, p-values, chance levels are summarised in Table 3.3. The results indicate a big difference between nested accuracy of 60% and non-nested accuracy of 90% in cross-validation methods. The average chance level for the nested cross-validation method is 53%, which is consistent with a 2-

<table>
<thead>
<tr>
<th>Reference</th>
<th>Data split</th>
<th>Feature Extraction</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Ind</td>
<td>WT, SE, CC, AR</td>
<td>SVM</td>
<td>95</td>
</tr>
<tr>
<td>7</td>
<td>Ind</td>
<td>Statistical, DWT</td>
<td>SVM</td>
<td>85</td>
</tr>
<tr>
<td>116</td>
<td>Ind</td>
<td>DWT</td>
<td>KNN</td>
<td>84</td>
</tr>
<tr>
<td>85</td>
<td>Both</td>
<td>SE</td>
<td>SVM</td>
<td>80</td>
</tr>
<tr>
<td>105</td>
<td>Dep</td>
<td>Statistical, MRMRM</td>
<td>SVM</td>
<td>73</td>
</tr>
<tr>
<td>84</td>
<td>Ind</td>
<td>AR</td>
<td>KNN</td>
<td>72</td>
</tr>
<tr>
<td>62</td>
<td>Ind</td>
<td>Statistical, PSD, HOC</td>
<td>KNN</td>
<td>70</td>
</tr>
<tr>
<td>83</td>
<td>Dep</td>
<td>FFT</td>
<td>NB</td>
<td>70</td>
</tr>
<tr>
<td>22</td>
<td>Ind</td>
<td>Statistical</td>
<td>C4.5</td>
<td>67</td>
</tr>
<tr>
<td>78</td>
<td>Dep</td>
<td>Statistical, HFD</td>
<td>SVM</td>
<td>63</td>
</tr>
<tr>
<td>111</td>
<td>Ind</td>
<td>HOSA</td>
<td>SVM</td>
<td>60</td>
</tr>
<tr>
<td>102</td>
<td>Ind</td>
<td>HOC, STFT</td>
<td>RF</td>
<td>40</td>
</tr>
</tbody>
</table>

Acronyms: Data split: subject-dependant (Dep), subject-independent (Ind). Features: auto-regressive (AR), cross correlation (CC), digital wavelet transform (DWT), Higuchi fractal dimension (HFD), higher order crossings (HOC), higher order spectral analysis (HOSA), maximum relevance minimum redundancy method (MRMRM), power spectral density (PSD), sample entropy (SE), short-time Fourier transform (STFT), wavelet transform (WT). Classifiers: Kth nearest neighbour (KNN), Naive Bayes (NB), random forest (RF), support vector machines (SVM).

class random chance level, whilst the average for the non-nested method is 67%. The chance level of 67% is similar to that reported by [9]. In addition, the accuracies of non-nested cross-validation methods up to 90% are consistent with those of previous studies claiming high accuracies despite not explaining their cross-validation methods thoroughly. The standard deviation of the two methods suggests there is no overlapping. The p-values of both the nested and non-nested cross-validation methods were statically significant.

3.2.5 Discussion

The literature survey results showed the average accuracy of all studies utilising NN classifiers to be 80%. These findings are consistent with the literature survey conducted by [11]. Table 3.4 shows all studies from literature survey [11] reporting classification accuracies using the same dataset. The different datasets were separated in an attempt to set a benchmark for the same particular dataset used in the literature survey conducted in this experiment: DEAP [12]. There were 12 studies; 6 reported accuracies less than or equal to 70%, 6 reported accuracies above 70%, and 1 reported an accuracy of 95%. However, these studies did not not include NNs, and therefore a new literature survey was needed.
Re-implementing the classification tests resulted in two accuracies of 60% and 90% corresponding to nested and non-nested cross-validation methods, respectively. Since the vast majority of studies have reported accuracies above 70%, it is now believed that most of these NN studies, where details have not been given, utilised non-nested cross-validation testing. This conclusion confirms the opinion of [2]. The reason NNs would perform poorly, if tested more appropriately, is that NNs require a large number of parameters to be optimised, and as a result, they need a large number of training trials [2], [64]. This is a general problem in BCI, where the number of trials is constrained by experiment length and cost. The same problem applies to SVM. An earlier version of this analysis was conducted to SVM classification in a smaller scale and was published in conference proceedings [65]. It was found that the use of non-nested cross-validation of $C$ and $\gamma$ parameters caused an increase of 13% in accuracy, applied to ERP-based BCI dataset. The study is attached as Appendix A.

The main contribution of this analysis is the quantification of the inflation in performance caused by non-nested cross-validated NN classification. This should motivate researchers to clearly describe their cross-validation methods and helps avoid confusion caused by seeing these variations in accuracy. In the literature survey discussed in Section 3.2.4, only two studies clearly stated the use of non-nested cross-validation. In the first study [9], the analysis included calculating chance level for the algorithm. The 2-class chance level of 63% was sufficient to illustrate the use of non-nested cross-validation with SVM classification, which was confirmed by these results. The second study [10] has provided an algorithm which indicated the model being trained and tested on the same data, which was replicated in this research and resulted in higher accuracies as discussed earlier. Many remaining studies still only describe their cross-validation methods vaguely, and therefore cause confusion. Due to the lack of ability to separate which cross-validation method was used, it is not possible to set an accurate state-of-the-art benchmark for each classifier. In this thesis, all classification tests are done with nested cross-validation, which has been shown to give the most robust and trustworthy estimates of performance and will not overstate the results. This will mean that the work is more repeatable and ultimately usable by others.
3.2.6 Summary

This section investigates performance inflation of NN algorithms tested on emotion recognition data DEAP. The aim was to establish the cause of the variation in accuracies by first conducting a literature survey and then re-implementing a NN classifier. A literature survey of emotion recognition algorithms showed the vast majority of studies testing NN classifiers reported accuracies above 80%. A critical analysis of cross-validation methods indicated that some researchers have utilised non-nested cross-validation methods, resulting in significantly higher performance and chance-level accuracies. To test this, the most NN classification with PSD features were re-implemented using both nested and non-nested cross-validation methods, and the results showed average performance and chance level accuracies of 60% and 53%, respectively, for nested cross-validation methods and 90% and 67%, respectively, for non-nested cross-validation methods. This section aimed to increase testing robustness by motivating researchers to clearly indicate their cross-validation methods. The following section compares various classifiers and show LR performs best for low number of training trials and EEG electrodes, tested on two datasets related this thesis. In this thesis, all subsequent accuracies are reported using nested cross-validation methods, because they are more reliable in generalised conditions.

3.3 Demonstrating logistic regression (LR) is the best performing classifier for low-channel count EEG in visual and emotion classification

3.3.1 Introduction

Section 3.2 discussed performance inflation caused by cross-validation in NN classification, and concluded that the use of nested cross-validation causes more analysis robustness. In this section, this method is going to be applied to find the best performing classifier in both visual and emotion recognitions.

As a new researcher, it is very difficult to choose the right classifier for robust EEG
analysis. Tens of classifier tests are published everyday, with every study claiming their algorithm is better than state-of-the-art. In this section, three classifiers that have widely been reported to perform well are tested on two relevant datasets with the aim of establishing the best classifier for this type of data. The literature is used to justify all algorithm development choices, including ERPs and frequency-domain feature extraction, spatial filtering, and classification. Covariances deployed with Riemannian geometry were chosen for feature extraction and spatial filtering methods in this algorithm. This is because Riemannian geometry-based feature extraction methods were found to outperform other state-of-the-art methods, earning first-place prizes in five international BCI competitions [66].

The algorithm described in this section was used in all relevant work in Chapters 4 and 5. However, this section contains information that is not repeated in significant detail elsewhere because the algorithm descriptions in other chapters focus on the pre-processing related to the specific setup and data collection. Section 3.3.2 describes the aim of the study. Sections 3.3.3, 3.3.4, and 3.3.5 provide the methods, results and discussion, respectively.

### 3.3.2 Aim

The aim is to test various classifiers and choose the best performing classifier on datasets with similar attributes to the work discussed in other chapters. The algorithm combinations are tested on different configurations representing a wide range of EEG experiments. Firstly, three classifiers were tested; LR, SVM, and NN. SVM and NN were chosen because of their common use in BCI work. LR was used because of its linearity and robustness with small number of training trials. Secondly, two datasets were chosen to cover ERP-based and emotion recognition BCIs. These two datasets have similarities with the data collected for Chapters 4 and 5, respectively. In addition, the two datasets are concerned with two types of BCI information; time-domain ERPs and frequency-domain free-running EEG, thus making the feature extraction generalisable for other type of BCIs such as P300-speller, SSVEP, and MI. Thirdly, the algorithm is tested on different channel counts such as 204 magnetoencephalography (MEG), 74 EEG, and 12 EEG channels. MEG is similar to EEG but it measures brain magnetic fields rather than electrical fields, which is the case in EEG [67]. The combination of EEG and
MEG could provide insight on the performance on various setups, and therefore enables choosing the right classifier for specific setups. Finally, the classifiers are also tested for a large number of training trials in the first dataset and a small number of training trials in the second dataset. This is important because some classifiers, such as MLP, are found to perform poorly with a small number of training trials, as discussed in Section 3.2. Therefore, validating the drawbacks of classifiers is useful for choosing the right one.

### 3.3.3 Methods

The analysis comprised several steps to achieve a system accuracy. These steps were explained and presented in Figure 2.1. Firstly, the data were obtained from public datasets, and the they were subsequently processed through feature extraction, spatial filtering, and vectorisation. Finally, the algorithm was tested with three classifiers to report accuracy. Two forms of features were used to extract free-running frequency and ERP information, to become useful for a wide range of BCI applications. These features were then combined as described in detail in the following sections.

### Datasets

There two datasets that have been used in this analysis.

**Openfmri00017 EEG and MEG face/non-face dataset:** The public dataset used for this analysis was obtained from [68]. It includes data from 16 participants. The experiment involved looking at a total of 554 pictures of two classes: faces and scrambled images. The data was recorded using a research-grade EEG/MEG device with a total of 204 MEG electrodes and 74 EEG electrodes at a sampling frequency of 500 Hz. This dataset has the same stimuli (same face/non-face pictures) as those used in Chapter 4, and it will be beneficial to this study in three ways. Firstly, it will help evaluate the developed algorithm by comparing its accuracy with those of previous studies that used the same dataset. This will ensure that the performance of the algorithm is satisfactory. Secondly, it will reduce variability between this study and the work in Chapter 4, as both studies use the same stimuli and set of electrodes with the only major difference being
EEG recorder quality. This allows for an accurate benchmarking of the work in Chapter 4. While the EEG recorder used in this dataset is research-grade and the one used in the Chapter 4 is commercial-grade. Thirdly, it will assist in comparing the 3 classifiers (LR, SVM, and MLP) and ranking their performances. In the analysis, three different configurations will be used to quantify the difference in accuracies between EEG channel selection: the use of data from 204 MEG electrodes, the use of data from 74 EEG electrodes, and the use of 12 EEG channels (AF3, F7, F3, FC5, P7, O1, O2, P8, FC6, F4, F8, AF4) as is similar to the Emotiv EPOC+ electrode positioning. This will be helpful in evaluating the new system’s performance employing the Emotiv used in Chapter 4.

**Dataset for emotion analysis using physiological signals (DEAP):** The dataset used for this analysis was obtained from [12], and it includes the results of 32 subjects, with equal male-female participation and aged 19-37 (mean 26.9), watching 40 one-minute music video clips that were categorised into different levels of arousal and valence emotive states. The clips were short-listed using a survey that contained 120 clips. A device with 32 electrodes at 512 Hz and another 12 peripheral channels was used to record physiological signals, which were not used in current analysis. Each subject rated the clips in terms of arousal, valence, and dominance on a continuous 1-9 scale. The pre-processed data (down-sampled to 128 Hz) were available publicly to download along with the participants’ ratings. The authors reported a classification accuracy of 62% for arousal and 57.6% for valence. The DEAP dataset was used because its setup highly resembles that of the experiment in Chapter 5. Both specify 32-channel research-grade EEG recorders, valence and arousal musical clips for stimuli, and a similar number of trials (40 for DEAP and 56 for Chapter 5).

**Frequency information extraction and spatial filtering**

The sample covariance matrix (SCM) of an input signal \( X \) of length \( N \) is defined by [69] as

\[
SCM = \frac{1}{N}XX^T. \tag{3.1}
\]

Covariance of frequency-domain feature was chosen for its ability to contrast the different classes and maximise information extraction with regards to spatial asym-
Subsequently, an advanced CSP method utilising Riemannian geometry, described by [66], [70], [71], was used. To apply Riemannian CSP spatial filtering, the mean of a single class’s covariances was needed. Given symmetric positive definite (SPD) matrices $P_1, P_2, \ldots, P_L$ the Riemannian mean $RM$, as defined in [71], is

$$RM(P_1, P_2, \ldots, P_L) = \arg \min_{P \in P(n)} \sum_{i=1}^{L} \delta^2_R(P, P_i),$$

where $\delta_R(P_1, P_2)$ is the Riemannian distance between two symmetric positive definite SPD matrices, $P_1$ and $P_2$ in $P(n)$, and is given by

$$\delta_R(P_1, P_2) = \left[ \sum_{l=1}^{L} \log^2 \frac{\lambda_l}{1 - \lambda_l} \right]^{1/2},$$

where $\lambda_l$ are the eigenvalues. Riemannian distance approximates Euclidean distance in the Euclidean feature space, which measures the separability of classes in this feature space. Next, the discriminative power of each spatial filter $w_j$ is related to its associated eigenvalue by the relationship

$$(A_1 + A_2)^{-1} A_1 w_j = \lambda_j w_j,$$

where $A_1$ and $A_2$ are the Riemannian mean matrices of classes 1 and 2, respectively. Thus, the filter was chosen based on what values maximise the quantity $|\lambda_j - 0.5|$. The filtered output $\hat{X}$ therefore was obtained by the direct projection of $SCM$ on each trial with the spatial filter $W$ that maximised the covariance between the classes:

$$\hat{X} = W^T SCM(X).$$

This method has shown improvements in classification accuracy over standard CSPs [71].

Riemannian geometry-based methods work equally well on all BCI paradigms, as they allow for the extraction of both time-domain (such as ERP-based visual BCIs) and frequency-domain features (such as motor imagery, emotion, and SSVEP). The only difference is how the matrices are mapped onto the tangent space. If both methods are combined, the paradigm and feature types do not have to be
known in advance. Moreover, Riemannian geometry allows for the classification of features in their matrix form, such as covariance matrices, instead of their vector form, which preserves the spatial information of adjacent-electrode data points and is therefore useful in BCI design. The Riemannian geometry-based method is robust to noise, and its improved generalisation capability is conditioned with a relatively small number of electrodes. Notably, its computational complexity increases cubically with the number of electrodes used. However, this can be avoided by spatial filtering, a very common technique that reduces the number of channels and increases the signal-to-noise ratio [2].

**ERP information extraction and spatial filtering**

This method, as described by [72], [73], calculated the average $P_k$ of the training trials $X_l, l = 1, 2, ..., L$, where $L$ is total number of training trials for each class $k$ separately:

$$P_k = \frac{1}{L} \sum_{l=1}^{L} X_l.$$  \hfill (3.6)

Thus, for each class $k$, the spatial filter $w$ that maximised the time-domain features of the class average $P_k$ over all training data trials $X$ follows a generalised Rayleigh quotient equation:

$$\hat{w}^* = \arg \max_w \left( \frac{w^T P_k P_k^T w}{w^T X X^T w} \right).$$  \hfill (3.7)

Therefore, the $m$ eigenvectors that correspond to the largest eigenvalues in the decomposition of $[P_k P_k^T (XX^T)^{-1}]$ are taken from each class to form $w_k$. As a result, the spatially filtered trial $Z_l$ was constructed by the linear projection of the spatial filter $W = [w_0, w_1]$ with the original trial data $X_l$:

$$Z_l = W^T X_l.$$  \hfill (3.8)

xDAWN good performance is due to its ability to maximise the separation between classes by increasing the signal-to-noise ratio and removing redundant data [2]. Since ERP-based time-domain features have smaller changes in amplitude compared to the background EEG, covariance matrices of a single trial cannot directly
extract them [74]. Therefore to use covariance matrices in time-domain analysis, a method by [74] was implemented; calculate the average of each class and then find the covariance between these averages and the trial data. This enabled the extraction of temporal information by comparing the energy of a single trial to the energy of the average ERPs. The covariance matrix indicated the similarity between the trial data and both classes. The following step, described in [72], extracted the SCM of each filtered trial \( \mathbf{Z}_l \) by concatenating it with the average \( \hat{\mathbf{P}}_k \) of each class \( k \):}

\[
\hat{\mathbf{Z}}_l = \begin{bmatrix}
\hat{\mathbf{P}}_0 \\
\hat{\mathbf{P}}_1 \\
\mathbf{Z}_l
\end{bmatrix}
\] (3.9)

Next, the SCM feature matrix of each trial was calculated as

\[
\mathbf{Z}_{SCM} = \frac{1}{N-1} \hat{\mathbf{Z}}_l \hat{\mathbf{Z}}_l^T,
\] (3.10)

where \( N \) is the number of samples in a single trial. This method is robust to noise, and outliers, such as data covariance, are calculated with an original, known ERP average [72]. This method has shown to be competent, as it won a classification competition of ERP-based BCI data [66].

**Multi-domain combination and vectorisation**

ERP and frequency features were combined to maximise information extraction. In addition, using both features allowed for the use of the same algorithm on different BCI systems. Moreover, it proved useful when the data was of an unknown BCI type. Combining time- and frequency-domain features to maximise information extraction, as opposed to relying on a single feature, leads to better performances [2]. However, this combination increased computational complexity and dimensionality and, as a result, it made offline processing slightly longer.

The SCM had a 2D form, whereas most classifiers accept features in a 1D vector form. To solve this, a simple vectorisation was done, as described by [69]. Concatenate each row in a matrix of size \( n \times n \) into one large column vector of length \( n(n + 1)/2 \) with the following modified half-vectorisation operator that has differ-
ent weights stacks on the upper-triangular part of $C$:

$$\text{vect}(C) = [C_{1,1}; \sqrt{2}C_{1,2}; C_{2,2}; \sqrt{2}C_{1,3}; \sqrt{2}C_{2,3}; C_{3,3}; \ldots; C_{n,n}]$$  \hspace{1cm} (3.11)$$

The $\sqrt{2}$ scalar was applied to all non-diagonal elements to preserve the equality of norms, i.e. $\|C\| = \|\text{vect}(C)\|$.

Classifiers

This subsection explains how and why the three classifiers LR, SVM, and NN have been utilised in this research. Each classifier’s main advantages and disadvantages will be described, and implementation details will be given, including an example algorithm for each methodology.

The first classifier was LR, which is a simple linear classifier that utilises statistical regression analysis to give a class probability if certain features or conditions are provided. LR was explained in Section 2.5.2. It was chosen because it can be trained with a small number of electrodes, which is the case with emotion BCI. LR does not assume linearity in relation to different variables [75]. In addition, it does not require Gaussian distributed independent variables [76]. A sample algorithm is provided in Algorithm 2.

### Algorithm 2 Logistic regression

1: procedure GETACCURACY
2:     \textit{cv} ← ShuffleSplit(n_splits=10)
3:     model\_ERP ← make\_pipeline(xDAWN, ERP\_covariance,vectorisor)
4:     model\_em ← make\_pipeline(covariance, CSP,vectorisor)
5:     model\_combined ← unite\_pipeline(model\_ERP, model\_em)
6:     model ← make\_pipeline(model\_combined, logistic\_regression)
7:     \textit{X} ← data (trials,channels,samples)
8:     \textit{y} ← labels
9:     accuracy ← cross\_validate(model, \textit{X}, \textit{y}, \textit{cv})
10:    return accuracy

The second tested classifier was SVM. Linear and non-linear SVMs are widely used and have been successful in different BCI setups [77]. As seen in Section 3.2, SVMs are the most common classifiers used in emotion BCIs. Moreover, SVMs have good generalisation properties due to their margin maximisation and regularisation properties [77]. In addition, SVMs are insensitive to overtraining and to the curse of dimensionality, where the number of training trials is relatively
small compared to the number of electrodes [77]. However, SVM computational complexity is slightly larger than linear, and the regularisation of simple classifiers such as LDA and LR, often helps with the generalisation caused by its outlying features. For example, regularised linear SVMs outperform unregularised LDA classifiers, and regularised non-linear radial basis function (RBF)-based SVMs outperform unregularised non-linear multilayer perceptron classifiers [77]. The aforementioned reasons encouraged implementing and benchmarking RBF-based SVM. However, RBF parameters conventionally known as $C$ and $\gamma$ require tuning for each specific set of features. Parameter $C$ represents the cost, expressly the classification surface smoothness to trade-off between the misclassification of training trials and a simpler decision surface, where $\gamma$ represents the impact of individual samples on the chosen support vectors. The cross-validation parameter tuning of $C$ ranged from 1 to 1000, and $\gamma$ ranged from 0.0001 to 0.1 with 10 folds each. A sample algorithm is provided in Algorithm 3.

**Algorithm 3 SVM**

<table>
<thead>
<tr>
<th>Line</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>procedure</strong> GetAccuracy</td>
</tr>
<tr>
<td>2</td>
<td>$cv \leftarrow$ ShuffleSplit(n_splits=10, test_ratio=0.2)</td>
</tr>
<tr>
<td>3</td>
<td>$model_ERP \leftarrow$ make_pipeline(xDAWN, ERP_covariance,vectorisor)</td>
</tr>
<tr>
<td>4</td>
<td>$model_em \leftarrow$ make_pipeline(covariance, CSP,vectorisor)</td>
</tr>
<tr>
<td>5</td>
<td>$model_combined \leftarrow$ unite_pipeline(model_ERP, model_em)</td>
</tr>
<tr>
<td>6</td>
<td>$model \leftarrow$ make_pipeline(model_combined, SVM_RBF)</td>
</tr>
<tr>
<td>7</td>
<td>$X \leftarrow$ data (trials,channels,samples)</td>
</tr>
<tr>
<td>8</td>
<td>$y \leftarrow$ labels</td>
</tr>
<tr>
<td>9</td>
<td>$model_GS \leftarrow$ GridSearchCV(model, params=$[\gamma=(0.0001-0.1), C=(1-1000)]$, CV)</td>
</tr>
<tr>
<td>10</td>
<td>$model_GS.fit(X_train,y_test)$</td>
</tr>
<tr>
<td>11</td>
<td>$accuracy \leftarrow$ cross_val_score(model_GS, X_test,y_test,cv)</td>
</tr>
<tr>
<td>12</td>
<td>return $accuracy$</td>
</tr>
</tbody>
</table>

The third classifier was based on neural networks (NNs). NNs have been described by [52] as one of the most appropriate tools to classify EEG data, as their natural neuron behaviour resembles that of the EEG behaviour produced by neurons. Moreover, NNs can derive meaning from complicated chaotic data, including EEG data [52]. On the other hand, many studies have reported that NNs need a large number of training trials to perform well, as previously seen in Section 3.2, which makes it unsuitable for many BCI applications. Some studies have shown that NNs were among the worst classifiers when tested data limited by a small number of training trials [2]. Although NNs are widely used in emotion classification studies, as seen in Section 3.2, there are still some difficulties with their performance expectations. Some studies have claimed accuracies up to 90%, whereas
others claim a mere 60%. Section 3.2 demonstrated how the use of PSD features with nested cross-validation methods can result in a low accuracy. However, as no previous studies have reported on the use of NNs with covariance matrices, this was included the scope of this analysis. As seen in Section 2.5.4, there are a few types of NNs such as MLP and CNN. The MLP classifier employed in this analysis had a sequential model. The first layer had an input dimension equal to the number of features in a single trial, with the same number of output neurons as number of features. The second layer with rectified linear units (RELU) activation had an output number of neurons equal to half of those in the previous layer, which would allow for enough information to be passed in from the previous layer. RELU activation was used to introduce non-linearity in the model, which makes it more practical for this application [10], [13]. The third layer followed a similar structure, and so on. The final layer had an output shape of one neuron to produce a 1 or 0 for either class. The loss was calculated using a binary cross-entropy metric, because of having two classes. A sample algorithm is provided in Algorithm 4.

Algorithm 4 NN

```python
1: procedure GetAccuracy
2:   cv ← ShuffleSplit(n_splits=10, test_ratio=0.2)
3:   model_ERP ← make_pipeline(xDAWN, ERP_covariance, vectorisor)
4:   model_em ← make_pipeline(covariance, CSP, vectorisor)
5:   model_combined ← unite_pipeline(model_ERP, model_em)
6:   model_MLP ← Sequential()
7:   model_MLP.add(Dense(1.0*n_feature), activation='relu')
8:   model_MLP.add(Dense(0.5*n_feature), activation='relu')
9:   model_MLP.add(Dense(0.25*n_feature), activation='relu')
10: ...
11:   model_MLP.add(Dense(1), activation='sigmoid'))
12:   model_MLP.compile(loss='binary_crossentropy', optimizer='rmsprop', metrics='auc'))
13:   model ← make_pipeline(model_combined, model_MLP)
14:   X ← data (trials, channels, samples)
15:   y ← labels
16:   accuracy ← cross_val_score(model, X, y, cv)
17:   return accuracy
```

In summary, LR is a linear classifier that is expected to perform well with a small number of training trials and is robust to noise. RBF-based SVM is a widespread efficient classifier that utilises a regularised non-linear kernel to separate classes. Its performance depends on two parameters $C$ and $\gamma$, which are tuned by cross-validation. NN-based classifiers have shown different performances across the literature because of different cross-validation methods. An MLP classifier was included for performance comparison purposes.
Cross-validation

As discussed thoroughly in Section 3.2, different cross-validation methods result in vastly different accuracies. In this test, a nested method was used to obtain the most precise accuracy values possible. Two cross-validation stages were completed; the first tuned classifier parameters, while the second tested classifier performance on new data. Each stage divided the data with a ratio of 20% for validation and testing. Each stage was repeated ten times to obtain an accurate performance. All accuracies are calculated using AUC, which was discussed in Section 2.5.7. The statistical significance was evaluated using the t-test, which was explained in Section 2.5.8.

In summary, Section 3.3.3 described the algorithm development process and provided justifications for each choice made. Two datasets were used because of their similarities to the work in this thesis. Features for ERP-based BCIs were spatially filtered to maximise the signal-to-noise ratio, and feature matrices were formed and then vectorised. On the other hand, features for frequency-domain BCIs, such as emotions, had covariance matrices extracted and were then spatially filtered with Riemannian CSPs before vectorisation. Both types of features were then combined and fed to the classifier for training and testing. Three classifiers were tested: a simple LR, the standard SVMs, and the advanced NNs. To avoid bias, a nested cross-validation method was used.

3.3.4 Results

Face/non-face dataset

The classification accuracies for all configurations, represented by AUC, are summarised in this section. First, Table 3.5 shows classification accuracies obtained using the 204 MEG electrodes. This indicates that the highest average accuracy was 88.44% for LR, whereas the lowest average accuracy was 78.52% for MLP. Moreover, the standard deviation for LR and SVM was less than 5% and for MLP was around 10%. All classifiers had statistically significant p-values less than 0.0001. The highest subject-dependent classification accuracy of 86.2% reported by [72] employed the same features in this analysis but used Riemannian geometry for
Table 3.5: 204 MEG classification accuracies (AUC) and p-values for LR, SVM, and NN

<table>
<thead>
<tr>
<th>Participant</th>
<th>LR (AUC)</th>
<th>SVM (AUC)</th>
<th>MLP (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>89</td>
<td>85</td>
<td>78</td>
</tr>
<tr>
<td>2</td>
<td>86</td>
<td>86</td>
<td>83</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>88</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
<td>83</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>85</td>
<td>82</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>83</td>
<td>80</td>
<td>75</td>
</tr>
<tr>
<td>7</td>
<td>92</td>
<td>91</td>
<td>83</td>
</tr>
<tr>
<td>8</td>
<td>91</td>
<td>91</td>
<td>87</td>
</tr>
<tr>
<td>9</td>
<td>86</td>
<td>85</td>
<td>45</td>
</tr>
<tr>
<td>10</td>
<td>88</td>
<td>88</td>
<td>86</td>
</tr>
<tr>
<td>11</td>
<td>82</td>
<td>81</td>
<td>72</td>
</tr>
<tr>
<td>12</td>
<td>90</td>
<td>88</td>
<td>78</td>
</tr>
<tr>
<td>13</td>
<td>89</td>
<td>82</td>
<td>81</td>
</tr>
<tr>
<td>14</td>
<td>94</td>
<td>93</td>
<td>85</td>
</tr>
<tr>
<td>15</td>
<td>95</td>
<td>94</td>
<td>90</td>
</tr>
<tr>
<td>16</td>
<td>88</td>
<td>86</td>
<td>78</td>
</tr>
</tbody>
</table>

Average: 88.44 (LR), 86.53 (SVM), 78.52 (MLP)
Standard deviation: 3.61 (LR), 4.44 (SVM), 10.85 (MLP)
P-value: 0.0001 (LR), 0.0001 (SVM), 0.0001 (MLP)

The use of SVM resulted in an accuracy of 86.5%, which was slightly better than the 86.2% that was obtained by [72]. This indicates that the chosen classification algorithm outperforms the current state-of-the-art algorithms.

The second configuration examines testing the algorithm using 74 EEG electrodes. Table 3.6 shows the results. The highest average accuracy was 93.7% for LR, whereas the lowest average accuracy was 87.5% for MLP. SVM outperformed MLP with an average accuracy of 90.44%.

Table 3.6: 74 EEG classification accuracies (AUC) and p-values for LR, SVM, and MLP

<table>
<thead>
<tr>
<th>Participant</th>
<th>LR (AUC)</th>
<th>SVM (AUC)</th>
<th>MLP (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94</td>
<td>87</td>
<td>92</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>87</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>85</td>
<td>78</td>
</tr>
<tr>
<td>4</td>
<td>96</td>
<td>94</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td>92</td>
<td>89</td>
<td>81</td>
</tr>
<tr>
<td>6</td>
<td>90</td>
<td>77</td>
<td>86</td>
</tr>
<tr>
<td>7</td>
<td>96</td>
<td>92</td>
<td>86</td>
</tr>
<tr>
<td>8</td>
<td>96</td>
<td>96</td>
<td>93</td>
</tr>
<tr>
<td>9</td>
<td>96</td>
<td>94</td>
<td>88</td>
</tr>
<tr>
<td>10</td>
<td>94</td>
<td>92</td>
<td>85</td>
</tr>
<tr>
<td>11</td>
<td>90</td>
<td>87</td>
<td>83</td>
</tr>
<tr>
<td>12</td>
<td>93</td>
<td>91</td>
<td>83</td>
</tr>
<tr>
<td>13</td>
<td>92</td>
<td>90</td>
<td>92</td>
</tr>
<tr>
<td>14</td>
<td>98</td>
<td>96</td>
<td>97</td>
</tr>
<tr>
<td>15</td>
<td>99</td>
<td>98</td>
<td>90</td>
</tr>
<tr>
<td>16</td>
<td>96</td>
<td>93</td>
<td>88</td>
</tr>
</tbody>
</table>

Average: 93.72 (LR), 90.44 (SVM), 87.53 (MLP)
Standard deviation: 3.18 (LR), 5.27 (SVM), 4.84 (MLP)
P-value: 0.0001 (LR), 0.0001 (SVM), 0.0001 (MLP)
Table 3.7: 12 EEG classification accuracies (AUC) and p-values for LR, SVM, and MLP

<table>
<thead>
<tr>
<th>Participant</th>
<th>LR (AUC)</th>
<th>SVM (AUC)</th>
<th>MLP (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78</td>
<td>71</td>
<td>74</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>66</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>68</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
<td>86</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>82</td>
<td>79</td>
<td>69</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>62</td>
<td>68</td>
</tr>
<tr>
<td>7</td>
<td>85</td>
<td>78</td>
<td>75</td>
</tr>
<tr>
<td>8</td>
<td>89</td>
<td>84</td>
<td>82</td>
</tr>
<tr>
<td>9</td>
<td>84</td>
<td>79</td>
<td>76</td>
</tr>
<tr>
<td>10</td>
<td>77</td>
<td>74</td>
<td>76</td>
</tr>
<tr>
<td>11</td>
<td>78</td>
<td>69</td>
<td>76</td>
</tr>
<tr>
<td>12</td>
<td>77</td>
<td>69</td>
<td>80</td>
</tr>
<tr>
<td>13</td>
<td>69</td>
<td>63</td>
<td>62</td>
</tr>
<tr>
<td>14</td>
<td>85</td>
<td>70</td>
<td>86</td>
</tr>
<tr>
<td>15</td>
<td>87</td>
<td>81</td>
<td>84</td>
</tr>
<tr>
<td>16</td>
<td>86</td>
<td>78</td>
<td>77</td>
</tr>
</tbody>
</table>

Average 80.21  73.61  75.66
Standard deviation 6.23  7.28  6.10
P-value 0.0001  0.0001  0.0001

accuracy of 90.4%. All classifiers had standard deviations around 5% and statistically significant p-values less than 0.0001. In addition, LR had the best classification performance compared with the wide-spread SVM and the state-of-the-art MLP. Comparing these results to a MEG classification competition [78] of faces versus scrambled images on the same dataset, the three best cross-subject classification accuracies were 75%, 73%, and 71%. The winner reported a subject-dependant classification accuracy of 86% and utilised ERP covariances as features and a vectorization in a tangent space to be classified with Riemannian geometry. The second-place competitor used down-sampled filtered raw data as features and logistic regression (LR) and random forest (RF) classifiers combined. The third-place competitor used support vector machines (SVMs). The leader board is available at [78]. A more recent study [79] found that the use of non-linear SVMs and a RF classifier with xDAWN spatial filtering resulted in subject-independent accuracies of 71% with EEG and 82% with MEG. These results indicate the competence of the classifier used in this analysis.

However, since the objective was to benchmark classification accuracy so that the system introduced in Chapter 4, the third configuration involved 12 EEG channels to match the number used by the commercial-grade Emotiv EPOC+. Table 3.7 shows the results of running the same algorithm with 12 electrodes. The highest average accuracy was 80.2% for LR, followed by 75.7% for MLP. The lowest av-
verage accuracy was 73.6% for SVM, where all classifiers had standard deviations between 6 and 7%. In addition, all classifiers had statistically significant p-values less than 0.0001. The accuracies of the 12 EEG channels were 13% less than those obtained using 74 EEG channels. This, as a result, gives insight onto the difference in accuracies caused by the number of EEG channels, in addition to the difference in accuracies caused by EEG and MEG technologies.

In conclusion, classification performance using LR with the MEG and EEG electrodes was superior to that of the SVM and MLP. The accuracies obtained for LR using 204 MEG, 74 EEG, and 12 EEG electrodes were 88%, 94%, and 80%, respectively.

**Emotion recognition dataset**

The classification accuracies for valence and arousal are summarised in Tables 3.8 and 3.9. Table 3.8 shows that the highest valence classification accuracy was 65.3% using LR. SVM and MLP had similar accuracies around 60%. All classifiers had similar standard deviations between 9% and 13% and had statistically significant p-values less than 0.0001. The use of covariance features with SVM and MLP is similar to the accuracy obtained by PSD features with MLP tested using nested cross-validation methods, as seen in Section 3.2. Table 3.9 shows that the highest arousal classification accuracy was 63.3% for LR. The lowest accuracy was 56.8% for SVM, whereas 61% was the accuracy of MLP. All classifiers had similar standard deviations between 11% and 12% and had statistically significant p-values less than 0.0001.

In conclusion, classification performance using LR for both valence and arousal was better than that of SVM and MLP. The accuracies obtained using LR for valence and arousal are 65.3% and 63.3%, respectively.

**3.3.5 Discussion**

The results showed LR outperforming SVM and MLP for both datasets. This is generally due to LR’s robustness to small training sizes [80], and low complexity allowing the classifier to avoid over-fitting [81]. In addition, LR have a less generalisation error than other classifiers and are easier to build [81]. Various studies
Table 3.8: Valence classification accuracies (AUC) and p-values for LR, SVM, and MLP

<table>
<thead>
<tr>
<th>Participant</th>
<th>LR (AUC)</th>
<th>SVM (AUC)</th>
<th>MLP (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69</td>
<td>56</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>54</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>68</td>
<td>64</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>47</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
<td>64</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>49</td>
<td>59</td>
<td>79</td>
</tr>
<tr>
<td>7</td>
<td>41</td>
<td>61</td>
<td>69</td>
</tr>
<tr>
<td>8</td>
<td>67</td>
<td>50</td>
<td>44</td>
</tr>
<tr>
<td>9</td>
<td>82</td>
<td>72</td>
<td>56</td>
</tr>
<tr>
<td>10</td>
<td>80</td>
<td>88</td>
<td>68</td>
</tr>
<tr>
<td>11</td>
<td>91</td>
<td>62</td>
<td>64</td>
</tr>
<tr>
<td>12</td>
<td>61</td>
<td>66</td>
<td>59</td>
</tr>
<tr>
<td>13</td>
<td>80</td>
<td>73</td>
<td>60</td>
</tr>
<tr>
<td>14</td>
<td>74</td>
<td>84</td>
<td>60</td>
</tr>
<tr>
<td>15</td>
<td>84</td>
<td>67</td>
<td>65</td>
</tr>
<tr>
<td>16</td>
<td>63</td>
<td>66</td>
<td>65</td>
</tr>
<tr>
<td>17</td>
<td>48</td>
<td>62</td>
<td>55</td>
</tr>
<tr>
<td>18</td>
<td>77</td>
<td>65</td>
<td>61</td>
</tr>
<tr>
<td>19</td>
<td>76</td>
<td>63</td>
<td>61</td>
</tr>
<tr>
<td>20</td>
<td>68</td>
<td>63</td>
<td>54</td>
</tr>
<tr>
<td>21</td>
<td>55</td>
<td>47</td>
<td>43</td>
</tr>
<tr>
<td>22</td>
<td>62</td>
<td>31</td>
<td>39</td>
</tr>
<tr>
<td>23</td>
<td>58</td>
<td>42</td>
<td>51</td>
</tr>
<tr>
<td>24</td>
<td>61</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td>25</td>
<td>77</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>26</td>
<td>66</td>
<td>52</td>
<td>54</td>
</tr>
<tr>
<td>27</td>
<td>68</td>
<td>65</td>
<td>79</td>
</tr>
<tr>
<td>28</td>
<td>76</td>
<td>53</td>
<td>58</td>
</tr>
<tr>
<td>29</td>
<td>70</td>
<td>82</td>
<td>69</td>
</tr>
<tr>
<td>30</td>
<td>65</td>
<td>48</td>
<td>70</td>
</tr>
<tr>
<td>31</td>
<td>42</td>
<td>57</td>
<td>46</td>
</tr>
<tr>
<td>32</td>
<td>70</td>
<td>59</td>
<td>56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Standard deviation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>65.29</td>
<td>13.30</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>60.71</td>
<td>12.07</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>59.45</td>
<td>9.47</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

have compared LR with other classifiers and have shown that its simple structure results in faster computation and often better results [82]–[84]. However, the fact that LR performed better could imply that the data used are linearly separable. Therefore, SVMs and NNs performances were affected by overfitting due to small number of training trials.

SVMs, on the other hand, performed worse than LR. Although most previous studies in emotion classification have utilised SVMs and reported accuracies up to 82%, as seen in Section 3.2, it is still believed that accuracy-reporting methods significantly affect these accuracies. The performance of MLP came last in most tests. This contradicts the results claimed by many, where different NN-based classifiers were reported to have accuracies between 80% and 90%, as seen in Section 3.2.
Table 3.9: Arousal classification accuracies (AUC) and p-values for LR, SVM, and MLP

<table>
<thead>
<tr>
<th>Participant</th>
<th>LR (AUC)</th>
<th>SVM (AUC)</th>
<th>MLP (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55</td>
<td>59</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>46</td>
<td>55</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>62</td>
<td>63</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>57</td>
<td>54</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>67</td>
<td>58</td>
<td>59</td>
</tr>
<tr>
<td>6</td>
<td>52</td>
<td>57</td>
<td>51</td>
</tr>
<tr>
<td>7</td>
<td>57</td>
<td>46</td>
<td>56</td>
</tr>
<tr>
<td>8</td>
<td>54</td>
<td>51</td>
<td>58</td>
</tr>
<tr>
<td>9</td>
<td>52</td>
<td>44</td>
<td>60</td>
</tr>
<tr>
<td>10</td>
<td>66</td>
<td>69</td>
<td>55</td>
</tr>
<tr>
<td>11</td>
<td>78</td>
<td>63</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>90</td>
<td>60</td>
<td>81</td>
</tr>
<tr>
<td>13</td>
<td>72</td>
<td>51</td>
<td>85</td>
</tr>
<tr>
<td>14</td>
<td>45</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>15</td>
<td>58</td>
<td>33</td>
<td>44</td>
</tr>
<tr>
<td>16</td>
<td>67</td>
<td>77</td>
<td>55</td>
</tr>
<tr>
<td>17</td>
<td>70</td>
<td>60</td>
<td>56</td>
</tr>
<tr>
<td>18</td>
<td>79</td>
<td>72</td>
<td>63</td>
</tr>
<tr>
<td>19</td>
<td>61</td>
<td>39</td>
<td>55</td>
</tr>
<tr>
<td>20</td>
<td>41</td>
<td>56</td>
<td>73</td>
</tr>
<tr>
<td>21</td>
<td>63</td>
<td>57</td>
<td>81</td>
</tr>
<tr>
<td>22</td>
<td>64</td>
<td>62</td>
<td>54</td>
</tr>
<tr>
<td>23</td>
<td>65</td>
<td>49</td>
<td>60</td>
</tr>
<tr>
<td>24</td>
<td>78</td>
<td>68</td>
<td>84</td>
</tr>
<tr>
<td>25</td>
<td>61</td>
<td>72</td>
<td>78</td>
</tr>
<tr>
<td>26</td>
<td>74</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>27</td>
<td>63</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>28</td>
<td>71</td>
<td>33</td>
<td>44</td>
</tr>
<tr>
<td>29</td>
<td>72</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>30</td>
<td>51</td>
<td>45</td>
<td>56</td>
</tr>
<tr>
<td>31</td>
<td>74</td>
<td>57</td>
<td>51</td>
</tr>
<tr>
<td>32</td>
<td>61</td>
<td>65</td>
<td>61</td>
</tr>
</tbody>
</table>

| Average     | 63.31    | 56.81     | 60.98     |
| Standard deviation | 11.00    | 11.18     | 12.16     |
| P-value   | 0.0001   | 0.0001    | 0.0001    |

Notably, the use of MLP in BCI has still not shown conclusive advancement when compared to other classifiers [2]. However, it was seen in Section 3.2 that NNs are sensitive to cross-validation methods. As a results, it is believed that many researchers perform NNs with non-nested cross-validation methods. With the use of nested cross-validation, LR has performed the best for both tests. Therefore, LR is going to be used in the remaining parts of the thesis.

3.3.6 Summary

Section 3.3 discussed the choice of classification algorithm by comparing LR to SVM and NNs. Two datasets relevant to this study were tested; an ERP-based
visual BCI dataset for comparison with data in Chapter 4 and a music-based emotion BCI dataset for comparison in Chapter 5. The algorithm had covariance matrices as features and LR, SVM, and NN as classifiers for comparison purposes. Overall, LR had the best performance on all tested datasets. For the ERP-based dataset, the accuracies of the 204 MEG, 74 EEG, and 12 EEG electrodes were 88%, 94%, and 80%, respectively. For the emotion dataset, the accuracies of valence and arousal were 65% and 63%, respectively.

3.4 Chapter summary

This chapter focussed on analysis robustness in designing a classification algorithm. This was achieved by investigating a problem causing variations on accuracies obtained by different cross-validation methods, discussed in Section 3.2. Once this was corrected, the best classifier was established by comparing LR, SVM, and NN, discussed in Section 3.3.

Section 3.2 investigated a problem that caused variations in accuracies in emotion BCIs. Some studies using NNs reported accuracies as high as 99% and others reported accuracies as low as 57%. It became clear that the employment of non-nested cross-validation methods caused significant increases in these accuracies. To confirm this, state-of-the-art MLP with PSD features were tested, and low accuracies resulted from nested cross-validation methods and high accuracies from non-nested cross-validation methods. As a result, MLP employed with non-nested cross-validation method is believed to cause inflated performance.

Next, Section 3.3 tested classifiers with state-of-the-art feature extraction and spatial filtering methods, tested on two datasets that were directly related to Chapters 4 and 5, respectively: an ERP-based visual BCI dataset with pictures of face and scrambled images as stimuli and a music-based emotion BCI dataset with musical video clips as stimuli. This included the use of SCMs in both ERP and frequency features. For the ERP features, the algorithm utilised xDAWN-based spatial filtering before constructing a matrix containing information about each class, which then was used to calculate a SCM. For the frequency-domain features, the SCM of each trial was calculated and then spatially filtered with an advanced Riemannian geometry-based CSP. Both time- and frequency-domain features were
then combined to maximise information extraction and create an algorithm suitable for all BCIs. Finally, three classifiers were tested: LR, SVM, and MLP. LR had the best performance followed by SVM and then MLP.

Once robustness has been established in this chapter, the next thing to look at is an application of ERP-based BCIs that aims to further improve system performance and user experience by tackling a user-fatigue problem caused by flickering lights.

References


Chapter 4

Improving user experience performance: A new non-flickering visual BCI paradigm

4.1 Introduction

Chapter 3 discussed robustness measures to improve performance in BCI. This chapter discusses user experience as a method of enhancing system performance. This work investigates the use of visual perception to improve user experience and reduce user fatigue with a commercial-grade EEG recorder.

The value of BCIs lies in providing positive experiences that deliver value for both patients and healthy users [1]. However, [2] found that a majority of BCI studies did not explicitly investigate user experience when designing their systems. Instead, most BCI systems are evaluated only on classification accuracy, but the user experience should be considered to understand if user needs are being adequately fulfilled [3]. However, there has been gradually increasing interest in improving user experience for BCI applications (e.g. see [4]–[7]). User experience could be assessed by testing usability factors such as ease of learning, user satisfaction with the system, and accessibility [1] and the results used to improve the BCI user experience. In this way, user acceptance and usability, in addition to classification accuracy could be enhanced [8].

BCI systems require continuous attention and working memory [9]. Such cognitive loads cause user fatigue and, thus, affects the effectiveness of communication [8].
This problem affects most common BCI paradigms, such as P300 speller, motor imagery (MI), and steady-state visually evoked potential (SSVEP). A study found that a continuous gaze at a P300 speller caused increased cognitive load and fatigue for users [10]. With MI, users must put forth an effort to sustain imagined movement of body parts for several seconds. This makes these systems difficult and tiring for users [8]. Furthermore, MI systems require training, and not everyone performs well with MI as compared to other modalities [11]. SSVEP, which is the focus of this study, exposes the user to a variety of continuously flickering lights, which is inconvenient, especially if the stimulus frequency lies between 5 and 30 Hz [12], [13]. Flickering lights enable communication in SSVEP by exposing the user to various sources of lights, flashing at different frequencies. Each source of light represents a unique command or communication message. For example, an SSVEP, with four direction arrows flashing at different frequencies, that controls a wheelchair manoeuvre direction. As a result, the user makes a choice by starting at the corresponding source of light, and the frequency of this light is detectable in the EEG. Furthermore, for users who have epilepsy, exposure to repeatedly flashing lights may increase the risk of seizures [12], [14], [15]. This is because the use of flashing frequencies much lower than the critical flicker frequency (CFF) induces user fatigue and thus affects system robustness [16]. To solve the problem of inconvenience associated with flickering lights, researchers [17] have suggested using frequencies above 30 Hz in SSVEP-based paradigms, at which point the flicker is no longer visually discriminable and it appears as a fixed light or video sequence, but resulted in decreased performance. In another study [18], the trade-off between classification accuracy and user experience was investigated in SSVEP. It was found that the use of frequency modulated SSVEPs, rather than traditional sinusoidal SSVEPs caused a decrease in classification accuracy but significantly improved user experience. In addition, another team of researchers [19] introduced an improved SSVEP paradigm by substituting the conventional light stimuli with chromatic green-blue higher frequency stimuli.

It is clear that a solution for the problem of the flickering lights is needed. This chapter contributes to that work by proposing the replacement of that source of user fatigue with a more relaxing source. By minimising user fatigue the user can be more relaxed and experience less stress, which in turn allows for a more consistent and reliable BCI system that can be used for a longer period of time. The
The proposed hypothesis is that exposure to non-flickering pictures will enable better user experience. Section 4.2 provides background by explaining how non-flickering images could be used to enable communication. Section 4.3 then describes the methods utilised to conduct the experiment, including a survey to test the hypothesis, an electronic design to enable accurate data collection, data preprocessing using ICA, and the classification algorithm. Sections 4.4 and 4.5 describe the outcomes of the experiment.

### 4.2 The face paradigm: Visual perception-based communication

The paradigm introduced in this chapter substitutes flickering lights with non-flickering pictures of different categories of visual perception to enable communication.

Visual perception in this context refers to the conceptual perception of pictures by looking at various images. Studies suggest that the brain behaves in a specific manner for different conceptual categories. For example, it has been shown by many researches that there are three common ERP components that are related to the face recognition process in visual perception. The first one is the N170 component at the posterior lateral areas of the brain (nearest electrode position, in the 10-10 system, is PO8), that is a negative amplitude ERP activated around 170 ms after stimulation [20]. Stimulation could be seeing a human face, or a dummy face [21]. Which indicates that what stimulates this ERP is the perception of a face and not a face. I.e. looking at a circle with some lines representing eyes and nose (known as Emojis), has the same perception and therefore the same brain behaviour. Other common ERPs are the N400 near the Cz electrode position, and the P600 near the Pz electrode position, which are activated during familiarity and recollection process, respectively [20]. Most of visual activity occurs at the visual cortex located in the occipital side of the brain as shown below. It has been shown in [22] that it is possible to identify an object category (animal or tool) from conceptual representations using picture-, text- or spoken-word stimuli with an accuracy of 89%. The study showed the main ERP components that were detected when stimulated by pictures are P1 at about 110 ms and N1 at about 160 ms.
Figure 4.1: The communication interface. (a) The software screenshot prior to stimulus. (b) The software screenshot during stimulus.
ms; these components were highest in energy at the posterior part of the head at infero-temporal and occipital electrodes. However, the study states the differences in ERPs between the animals and the tools categories are in the morphology of the signals, where the P1 component peaked earlier for animals than tools. Whereas the N1 component had larger energy for animals than tools. Also, at frontal electrodes the signal showed more negativity for animals than tools for the broad components that lasted from about 280 to 550 ms [22]. Another study [23] showed that conceptual imagery classification of images of a goat or a fire may be used to achieve an accuracy of over 70%. In this new paradigm, this brain behaviour of visual categories is used to enable communication. In particular, pictures of faces against scrambled pictures to express choices instead of continuously staring at flickering lights. The brain behaviour, separated by the features of N170 ERPs, is exploited to classify the choices.

Participants were asked to answer a list of questions by staring at pictures, and the EEG behaviour determined by their visual perception was used to determine their choice. The software started each question by presenting a message window that asked simple yes and no questions. The answer screen would then appear with the options for YES on the left and NO on the right. The participant answered the question by staring at the cross sign beneath either YES and NO according to their choice (see Figure 4.1a). A timer was represented by a growing red bar at the top of the screen, and when the timer ran out, two random images appeared where cross signs had been. One image was always of a face and the other was a scrambled picture, both of which were randomly chosen by the software. Further, the association of a face with a yes or no was randomised to prevent participants predicting a pattern. The answers were identified by classification of the EEG reading as either face or non-face (scrambled). It was then confirmed by cross-checking the recorded images shown at yes or no box. For example, if a picture was classified as a face, then the system checked the record to determine whether a face appeared for yes or no, and so on. Each picture appeared for 500 ms and then they were replaced by pictures of circles (see Figure 4.1b). The pictures were obtained from the dataset [24], and identical procedures were followed to match their dataset. This enabled comparison on the EEG data obtained in this experiment using a commercial-grade EEG, and the data obtained from [24], which used a research-grade EEG.
Table 4.1: List of 50 questions used in the paradigm. Questions were presented twice to enable 100 trials

<table>
<thead>
<tr>
<th>Question</th>
<th>Trivial answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can ducks quack?</td>
<td>Yes</td>
</tr>
<tr>
<td>Are plastic bottles made from wood?</td>
<td>No</td>
</tr>
<tr>
<td>Do fish eat ice cream?</td>
<td>No</td>
</tr>
<tr>
<td>Can fish swim?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is the opposite of black, white?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is 1 metre larger than 1 millimetre?</td>
<td>Yes</td>
</tr>
<tr>
<td>Does bread bleed?</td>
<td>No</td>
</tr>
<tr>
<td>Does the sun emit light?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is the moon green?</td>
<td>No</td>
</tr>
<tr>
<td>Can rocks speak English?</td>
<td>No</td>
</tr>
<tr>
<td>Can knives be sharp?</td>
<td>Yes</td>
</tr>
<tr>
<td>Does 1+1 equal 5?</td>
<td>No</td>
</tr>
<tr>
<td>Can trees run 10 mph (miles per hour)?</td>
<td>No</td>
</tr>
<tr>
<td>Do humans need to breathe?</td>
<td>Yes</td>
</tr>
<tr>
<td>Can someone fit a building into their trousers pocket?</td>
<td>No</td>
</tr>
<tr>
<td>Can cows produce milk?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do lamps need energy to run?</td>
<td>Yes</td>
</tr>
<tr>
<td>Does the ocean have water?</td>
<td>Yes</td>
</tr>
<tr>
<td>Can horses run?</td>
<td>Yes</td>
</tr>
<tr>
<td>Can cows produce chocolate milk?</td>
<td>No</td>
</tr>
<tr>
<td>Is the letter (S) a vowel?</td>
<td>No</td>
</tr>
<tr>
<td>Can scissors cut?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is Planet Pluto a 5-min walk from here?</td>
<td>No</td>
</tr>
<tr>
<td>Is the earth cube shaped?</td>
<td>No</td>
</tr>
<tr>
<td>Does fire feed on water?</td>
<td>No</td>
</tr>
<tr>
<td>Does 2+2 equal 17?</td>
<td>No</td>
</tr>
<tr>
<td>Can dogs bark?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do cars need fuel to operate?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do dinosaurs still exist?</td>
<td>No</td>
</tr>
<tr>
<td>Can flies bite?</td>
<td>Yes</td>
</tr>
<tr>
<td>Are oranges orange?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is the letter (A) a vowel?</td>
<td>Yes</td>
</tr>
<tr>
<td>Can dishes talk?</td>
<td>No</td>
</tr>
<tr>
<td>Can biscuits walk?</td>
<td>No</td>
</tr>
<tr>
<td>Can animals fly airplanes?</td>
<td>No</td>
</tr>
<tr>
<td>Can infants drive cars?</td>
<td>No</td>
</tr>
<tr>
<td>Is infinity larger than 10?</td>
<td>Yes</td>
</tr>
<tr>
<td>Can screwdrivers drive screws?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do vegetarians eat meat?</td>
<td>No</td>
</tr>
<tr>
<td>Can printers print?</td>
<td>Yes</td>
</tr>
<tr>
<td>Can humans walk?</td>
<td>Yes</td>
</tr>
<tr>
<td>Can camels use smartphones?</td>
<td>No</td>
</tr>
<tr>
<td>Does soap clean stuff?</td>
<td>Yes</td>
</tr>
<tr>
<td>Can a calculator find 1+1 value?</td>
<td>Yes</td>
</tr>
<tr>
<td>Are infants born with algebra knowledge?</td>
<td>No</td>
</tr>
<tr>
<td>Is the sky dark in a clear sunny day?</td>
<td>No</td>
</tr>
<tr>
<td>Is the opposite of up; down?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is 1 kilogram larger than 1 gram?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do cars have feelings?</td>
<td>No</td>
</tr>
<tr>
<td>Is the sun smaller than a peanut?</td>
<td>No</td>
</tr>
</tbody>
</table>
A few factors were considered to ensure the effectiveness during the application. Firstly, a set of 100 questions were defined to ensure an equal ratio of yes and no questions. Secondly, the questions were simple, factual, and not based on personal experiences, and therefore they had predictable answers. The list questions is presented in Table 4.1. Thirdly, a questionnaire-based survey was conducted to validate the proposed value. This survey is described in Section 4.3.2. Finally, design and implementation of a low-cost synchronisation circuit was needed to enable exact timing of stimulus onsets, which is discussed in Section 4.3.3. In the experiment, there were ten volunteers participating; the experiments were 30 to 45 minutes in length, with a setup time of less than 5 minutes for each participant. Using the 10-20 electrode location system, the electrode locations were AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4 with reference points at P3 and P4. Saline was used to wet all electrodes. For reference electrodes, an odour-free, colour-free, water-soluble gel was used to ensure conductivity if the saline dried out.

4.3 Methods

4.3.1 Introduction

This section explains the methods undertaken to test the face paradigm. Section 4.3.2 explains the questionnaire used to test the hypothesis that the face paradigm would provide a better user experience. Section 4.3.3 discusses an electronic contribution that was required for making a device to enable synchronisation of the EEG recorder. Section 4.3.4 describes a particular problem with data noise associated with the commercial-grade EEG that necessitated the involvement of ICA for preprocessing. Lastly, Section 4.3.5 describes the algorithm used for classification.

4.3.2 User experience survey

To assess the paradigm’s proposed value of convenience, a survey was conducted to record opinions of the face paradigm and the SSVEP (with flickering images) method. The experiment prompted the user to choose a paradigm (without stating the names of the paradigms or the technology behind them; see Figure 4.2a).
Figure 4.2: User experience experiment setup. (a) The starting point of the survey test. The participant was asked which paradigm (anonymised) they wanted to try first. (b) A screenshot of SSVEP answering with two options. Face paradigm is shown in Figure 4.1b.
A number of questions then appeared, and participants answered using either the face paradigm described above or SSVEP by looking at two flickering boxes on the screen as seen in Figure 4.2b, with the words YES and NO above them. The same questions were used for both paradigms. When the participants finished both paradigms (no EEG recorded), they were asked to complete a questionnaire that asked about relaxation, boredom, and ease of use. Fifty volunteers participated in the survey (34 females, mean age 22.86), and the results are discussed in Section 4.4.1. The following section describes a technical contribution that was required in order to synchronise the data.

### 4.3.3 Wirelessly synchronised commercial-grade EEG

The device used in this experiment was a low-cost Emotiv EPOC+ that was chosen because it takes only a short time to prepare and, therefore, improved the user experience. However, as discussed in Section 2.2.3, the Emotiv lacks a synchronisation peripheral, which makes it difficult to detect ERP information and, therefore, negatively impacted the classification performance. To solve this problem, a low-cost device was designed and implemented to enable synchronisation. The merit in this work is the design of a low-cost device to perform similarly to research-grade EEG, and with simple, yet effective, electronic engineering implementation (see Figure 4.3).

A synchronisation circuit was required to allow precise timing of the triggers sent to the Emotiv when the images should appear on the screen. The lack of a channel dedicated to receive trigger signals necessitated using two original EEG channels to allow receipt of signals. The channels selected were T7 and T8 because of their distant locations from visual areas of interest. The overall system is shown in Figure 4.4.

A battery-based system was developed to provide a direct link to the electrodes from the stimulus software. The software transmitted a signal via serial communication to a USB-connected microcontroller (MCU). The MCU had a photodiode attached to the monitor to detect image appearance. It transmitted a 433-MHz radio frequency (RF) signal to the receiver attached to the EEG recording machine, as shown in Figure 4.4. The schematic diagram of the circuit is shown in
Figure 4.3: The setup for the system. The EEG recording machine along with the synchronisation circuit inside a 3D printed enclosure. The thickness of the walls is less than 1 mm for a relatively light weight that did not affect the balance of the headset. The 3D model was designed with a small hanger that allowed it to be attached to the Emotiv without the need for screws or glueing.
Figure 4.4: An overview of the synchronisation device attached to the EEG recorder. The system on the left is the EEG recorder and the RF receiving device, powered by 3 LR44 batteries. The system on the right is the microcontroller (MCU) to transmit the RF signal.

Figure 4.5. The receiving device which was enabled with RF communication triggered one of the electrodes of the EEG machine (T8) with a number of 330 µV pulses for 8 ms, equivalent to 1 sample with a sampling frequency of 128 Hz (see Figure 4.6). The ground of the circuit was connected to another electrode (T7), both electrodes were biased to the reference electrode using 500 kΩ resistors. In case of a failure, the resistors would limit any current flow to 9 µA. This limit was less than 10 µA for CF Applied Part according to IEC 60601-1 requirements. Previous work conducted by others [25], [26] utilised infrared communication, which was limited by a distance of 1 meter between the receiver and the transmitter and required the receiver to be facing the transmitter for the signals to be recognisable. The alternative to this solution was to use a Bluetooth transmission. However, Bluetooth involved pairing two devices and an automatic lost packet retransmission protocol [27], and therefore, added variability in timing while transmitting. In this work, a simpler, novel form of RF communication was employed, which broadcast the signals, with numbered labels, to account for any lost packages. The RF transmit and receive modules had a couple of features. Firstly, broadcasting messages required no pairing, to minimise variant delays. It had a fixed transmission time, measured, using a serial monitor, at 17 ms per signal. This is equivalent to 2 timestamps at a sampling rate of 128 Hz. Secondly, as with any wire-
Figure 4.5: Circuit diagram for the synchronisation device attached to the EEG recorder. The EEG recorder electrodes T7, T8 and drive right leg (DRL) are connected to the microcontroller (MCU). The voltage was reduced using voltage divider from 3.3V to 330 µV; 500 kΩ resistors were used for safety to limit current. The MCU was an Atmega328p, known as Arduino, powered at 3.3V. Signals were received using 433-MHz RF modules.
Figure 4.6: Synchronisation triggers showing the data by subtracting EEG data from electrodes T7 and T8. Each pulse represents an additional delay of 17 ms. Number of pulses varied from 3 to 10.

less mode of communication, it is to be expected that a number of packets will be lost. Thus, each time the transmitter attempted to send a packet, it labelled the packet with a number. Therefore, the receiver knew how many packets were lost and, given their known delay times of 17 ms each, could determine the exact time of the onset. The receiver then sent a fixed number of pulses to the EEG channels indicating the time. Figure 4.6 shows samples of data from the EEG channels dedicated for receiving the triggers. The number of pulses shown indicate the number of milliseconds since the image appeared on the screen. Once the signals were synchronised, the data had to be filtered to remove noises using ICA as discussed next.
4.3.4 ICA data preprocessing

The raw EEG data required a number of preprocessing operations before feature extraction. Firstly, undesired frequency components were removed by applying a band pass filter with cut-off frequencies of 1 and 20 Hz. This was important for removal of low frequencies such as offsets and high frequencies such as mains noise at 50 Hz. This range was chosen because it maintained ERP information [28].

The second step involved the use of ‘FastICA’ [29] to compute 12 independent components corresponding to the 12 recording channels; ICA was explained in Section 2.4.2. ICA was used here because of its ability to find independence in the signal, unlike whitening that only solves for uncorrelatedness, which is weaker than independence as it only solves for half of the number of components of an arbitrary matrix [30]. In addition, ICA is superior to principle component analysis (PCA), as ICA is capable of finding the original spatial information of the signals unlike PCA [30]. This is a very important feature that is directly related to this specific application. Conserving spatial information is critical because of the relations between the EEG activities in various areas of the brain separating mental states. For example, motor imagery (MI) right and left hand movements are solely separated by their right and left side hemisphere brain activities. As a result, conserving spatial information by using ICA is key. In general, ICA requires a large number of electrodes (30+) to work effectively [31]. In this case, it was a satisfactory solution because there were only a small number of independent components of interest to retain, specifically, those that are associated with visible ERPs, either at the front of the brain, with positive potential (given the reference point was at P3), or negative potentials at the occipital part of the brain. The remaining components were removed. Figure 4.7 shows an example of one participant’s ICA data. Components are ordered by spectral power, where component 0 was the strongest, and component 11 was the weakest. Typically, the first couple of components are noises with large amplitudes. The noise derives from various sources, including electromyography (EMG) muscle noise, eye movement and gross body motion causing electromagnetic currents flowing through the EEG recorder wires [32].

An example of the process for selecting which components to remove and which to keep is illustrated in Figure 4.8. The figure shows the decision-making process for
three components. The epochs image and ERP box for component 1 shows a lack of ERP components. In addition, the source of the power is not localised in any specific area but, rather, distributed, unlike ERP features. This indicates that the component was noise and its removal was required. The epochs image and ERP boxes for components 2 and 4 show clear ERP peaks and troughs. Component 2 was localised in the back of the brain at visual (occipital) cortex. A negative ERP is consistent with all epochs and their average shows a clear N170 trough. This component was also informative and was not removed. Component 4 was localised at the front of the brain. A positive ERP is consistent with all epochs, and their average at the bottom shows a clear P300 peak. Thus, this component was informative and was not removed. The same process was repeated for all components and all participant data. The removal of all noisy components resulted in much clearer EEG data, as seen in Figure 4.9. The figure shows an example of trials obtained from averaging a single participant’s data, with peaks and troughs representing ERPs. Electrodes O2 and P8 show the clearest visible N170 and P300 ERPs because of their location in the visual cortex and having minimal noise. Other electrodes, such as FC5, P7, FC6, and AF3 have clearer ERPs after noise removal. On the other hand, a few electrodes, including AF4 and F4, lost informative ERPs in the process, which is a drawback of using fewer electrodes with this method.
Figure 4.8: An example of the decision-making process for three ICA components.

The boxes on the right represent the epoch images for the 100 epochs concatenated vertically, with the average epoch and its standard deviation plotted below them. The dotted purple line shows the time of recording the onset. The ERPs appeared 1.5 s later than expected due to different delays in synchronisation. These delays were consistent with all participants and, therefore, could be accounted for; they were caused by the delay in presenting the images to the screen. The red circles show ERP components found in corresponding components. The head images on the left indicate the location of the source of signals, where red indicates positive voltages and blue indicates negative. The decision is based on success in observation of ERP information.
Once the noise components were removed, the data were used for feature extraction, spatial filtering, and classification, as discussed next.

### 4.3.5 xDAWN spatial filtering and LR classification

The classification algorithm consisted of three main parts; spatial filtering, feature extraction, and classification (see Figure 4.10). Firstly, xDAWN spatial filtering, as described by [33] and [34], was used to calculate the average $P_k$ of training trials $X_l, l = 1, 2, ..., L$, where $L$ is total number of training trials for each class $k$ separately:

$$P_k = \frac{1}{L} \sum_{l=1}^{L} X_l. \quad (4.1)$$

Thus, for each class $k$, the spatial filter $w$ that maximised the time-domain features of the class average $P_k$ over all training data trials $X$ followed a generalised
Rayleigh quotient equation:

$$\hat{w}^* = \arg \max_w \left( \frac{w^T P_k P_k^T w}{w^T X X^T w} \right). \quad (4.2)$$

Therefore, the $m$ eigenvectors that corresponded to the largest eigenvalues in the decomposition of $[P_k P_k^T (X X^T)^{-1}]$ were taken from each class to form $w_k$. Form this, the spatially filtered trial $Z_l$ was constructed by the linear projection of the spatial filter $W = [w_0, w_1]$ with the original trial data $X_l$:

$$Z_l = W^T X_l. \quad (4.3)$$

xDAWN was chosen due to its ability to maximise the separation between classes by increasing the signal-to-noise ratio and removing redundant data [35]. Since ERP-based time-domain features have smaller changes in amplitude as compared to the background EEG, covariance matrices of a single trial cannot directly extract them [36]. Therefore, to use covariance matrices in time-domain analysis, a method by [36] was implemented: calculate the average of each class and then find the covariance between these averages and the trial data. This enabled the extraction of temporal information by comparing the energy of a single trial to the energy of the average ERPs. The covariance matrix indicated the similarity between the trial data and both classes. Secondly, feature extraction, as described by [33], calculated the sample covariance matrix (SCM) of each filtered trial $Z_l$ by concatenating it with the average $\hat{P}_k$ of each class $k$:

$$\hat{Z}_l = \begin{bmatrix} \hat{P}_0 \\ \hat{P}_1 \\ Z_l \end{bmatrix} \quad (4.4)$$

Next, the SCM feature matrix of each trial was calculated as

$$Z_{SCM} = \frac{1}{N-1} \hat{Z}_l \hat{Z}_l^T, \quad (4.5)$$

where $N$ is the number of samples in a single trial. This method is robust to noise, and outliers, such as data covariance are calculated with an original, known ERP average [33]. This was followed by vectorisation of the matrices, which was de-
Finally, a LR classifier was used in this analysis, which is a linear classifier that utilises statistical regression analysis to give a class probability if certain features and/or conditions are provided. LR was explained in Section 2.5.2. It was chosen because it can be trained with a small number of electrodes, which is the case with emotion BCI. LR does not assume linearity in relation to different variables [37]. In addition, it does not require Gaussian distributed independent variables [38]. For this type of data, LR was found to be the best performing classifier, as illustrated in Section 3.3. A sample algorithm is provided in Algorithm 5.

As discussed in Section 3.2, different cross-validation methods result in vastly different accuracies. In this analysis, a nested cross-validation method was used to
obtain accurate performance indicators. In addition, all accuracies were calculated using AUC and verified for statistical significance using the t-test, which were explained in Sections 2.5.7 and 2.5.8, respectively. Next, a summary of methods is presented.

4.3.6 Summary

This section described the methods used to enable the face paradigm. Firstly, the user experience survey asked questions about relaxation, boredom, and ease of use with regards to BCI. Secondly, RF transmitting and receiving devices were designed and implemented to enable accurate triggering to the Emotiv. The triggers were tested to successfully receive pulses at the EEG channels. This achievement enabled the use of a commercial-grade EEG in a setup that would typically require research-grade EEG. Thirdly, the specific preprocessing necessitated by the quality of the data was accomplished with the use of ICA to remove eye movement and other artefacts. Finally, an algorithm comprising xDAWN and covariance matrices was used with LR classification to enable testing the data. The results of the survey and classification accuracies are described next.

4.4 Results

The results of the survey and classification accuracies are presented in this section.
4.4.1 User experience survey

The results of the questionnaire are presented in Table 4.2, which shows that 88% of participants found the face paradigm more relaxing to use due to the reduced exposure to flickering lights associated with SSVEP decision-making. It also shows that 36% of the participants found the face paradigm more boring to use, and 46% thought it was easier to use. However, to confirm the statistical significance of the results, a t-test was conducted, taking each participant answer to the face paradigm as value 1 and to SSVEP as 0 to calculate the p-value. Table 4.2 shows that the p-values for the relaxation and boredom questions were less than 0.05, and thus were statistically significant. The results support this being a new paradigm with an improved user experience. On the other hand, the ease of use p-value of 0.5768 determined the results were not statistically significant. Therefore, the survey could not conclude which paradigm was easier to use.

This was an important test for validating the concept of convenience in using non-flickering images. The remaining part of the analysis was to enable communication with an acceptable performance. However, there were a couple of limitations with the survey. Firstly, Each question could be answered only with one of two choice, which could have implied to participants that one choice was better than the other. A future survey should use a continuous scale, which yields better results [39]. Secondly, although the paradigms were anonymous for the user to choose from, the order of the paradigms was fixed (i.e. face paradigm always number 1). This lack of control over the order of the paradigms may influenced the results causing the first paradigm to always feel less exhausting due to the fact of always appearing first.

4.4.2 Classification results

Table 4.3 shows the accuracies obtained for face and non-face classification. The results show the average accuracy was 67.14%. The highest participant accuracy
Table 4.3: Face versus non-face classification accuracies; 10 participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Accuracy (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
</tr>
<tr>
<td>5</td>
<td>71</td>
</tr>
<tr>
<td>6</td>
<td>74</td>
</tr>
<tr>
<td>7</td>
<td>55</td>
</tr>
<tr>
<td>8</td>
<td>66</td>
</tr>
<tr>
<td>9</td>
<td>65</td>
</tr>
<tr>
<td>10</td>
<td>66</td>
</tr>
</tbody>
</table>

Average 67.14
P-value 0.0001

was 78% and the lowest was 55%. Four participants scored over 70% and two scored below 60%. The p-value was less than 0.0001, which is statistically significant.

4.5 Discussion

This chapter presented a system that could potentially replace BCI paradigms based on flickering lights. The system utilised brain behaviour associated with face image perception to enable communication. It was applied using a commercial-grade EEG that featured convenient setup.

The survey results indicated that 88% of the participants found the face paradigm to be more relaxing to use than SSVEP. This proposed enhancement in user experience is meant to reduce user fatigue caused by the flickering lights in SSVEP. BCI systems are generally evaluated based on classification accuracies and information transfer rates, in order for this technology to have real-world applicability, user experience must be studied [40], which is what was achieved in this chapter. Moreover, BCI user experience is considered a part of system evaluation [1], and usability measures such as enjoyment are also important in user experience [1], [41].

The classification accuracy of the system was 67%, but two factors must be looked at to benchmark this accuracy. Firstly, the performance of algorithm used here was discussed in Chapter 3 with regard to a test on a public dataset. The classification accuracy of the algorithm was found to be competent as discussed in Chapter 3. Secondly, the quality of data obtained from the commercial-grade EEG and the number of channels also affect classification accuracy. Research-grade EEG
recorders perform significantly better than commercial-grade models. One study tested the classification accuracy of an Emotiv EPOC (commercial-grade EEG) with 14 electrodes against that of a Biosemi headset (research-grade EEG) with 32 electrodes using the oddball paradigm [42]. The study found that the Biosemi headset’s performance was 88.5% and the Emotiv EPOC’s was 61.7%. Many other studies have examined the performance of commercial-grade EEG recorders, with other visually evoked BCIs, like SSVEP, and obtained similar results [43], [44].

The effect of the number of channels was discussed in Chapter 3. It was found that 74 EEG electrodes resulted in an accuracy of 94%, while reducing the number of electrodes to the same 12 used in this experiment resulted in 80% accuracy. In addition, it was found that while current commercial-grade EEG data quality affects classification accuracy, with future hardware development for commercial EEG, performance would be better and user fatigue minimised. As a result, these findings suggest that there is no theoretical barrier for this paradigm to achieve high accuracies if the appropriate equipment is used.

The information Transfer Rate (ITR) in bits/second as described by [45] is defined as

\[
ITR = \left[ \log_2 N + P \log_2 P + (1 - P) \log_2 \frac{1 - P}{N - 1} \right] \frac{60}{S},
\]

where \(N\) is the number of different classes, \(P\) is the accuracy of classification, and \(S\) is the length of a single trial. I.e. \(\frac{60}{S}\) is the number of trials per minute. The ITR of the current face paradigm, using two classes, a commercial-grade EEG accuracy of 67%, and a trial length of 1 second, was found to be 5.1 bits/second. This is much lower than the state-of-the-art SSVEP rate at 62 bits/second (e.g. see [46]). The ITR is proportionally related to the classification accuracy and the number of classes. Firstly, the current average accuracy of 67% has limited the ITR to 5.1 bits/second. However, the accuracy of 94%, obtained with research-grade EEG in Chapter 3, could increase the ITR to 40.3 bits/second, which is eight times higher than the ITR obtained with the commercial-grade EEG, but less than SSVEP. Secondly, the number of classes in the current version of the paradigm is limited by two classes. This is clearly a disadvantage when compared to SSVEP that could have more than three classes. Increasing the number of classes to three could be accomplished by having different categories. For example, a classification of familiar face, non-familiar face, and non-face pictures was achieved.
by [24]. In addition, another study has investigated a stimulation with pictures of six different facial expressions: anger, disgust, fear, neutrality, sadness, and happiness [47]. This could potentially be used to increase the number of classes. However, increasing the number of classes with different visual perceptions might degrade classification accuracy. As a result, the face paradigm has a limiting disadvantage of low ITR.

The proposed system featured an improved user experience, but decreased classification accuracy and ITR. The survey concluded that 88% of participants found looking at faces and scrambled images more relaxing than looking at flickering images for communication. However, the accuracy when tested using commercial-grade EEG was found to be 67%, with an ITR of 5.1 bits/second. A trade-off between convenience and ITR is necessary in this case. As a result, flickering images could potentially be replaced by visual perception for BCI users where convenience is more important than ITR. This

This chapter aimed to shed light on the problem of user fatigue and propose a solution that would represent a step towards improving BCI user experience. Significant progress has been made towards BCI classification algorithm development, but user experience has not been examined enough. It is important to consider user experience, as it is what converts technologies to commercial application. In addition, increased user fatigue eventually affects attention and system performance. The system showed convenience but also had some degradation in accuracy due to the use of commercial-grade EEG. In addition to attempting a first-step solution for improving BCI user experience, this research was meant to be an inspiration for researchers to study user experience in BCI applications. EEG emotion and cognitive load recognitions could be used as quantitative-objective measures of user experience [3].

The work conducted in Chapter 5 discusses emotions recognition in music. The same technology could be used to give insight and improve on user experience.
4.6 Chapter summary

The aim of this chapter was to improve user experience through the creation of a new paradigm based on visual perception in order to eliminate the inconvenience of flickering lights. The face paradigm involved looking at non-flickering pictures of faces and scrambled images to enable communication.

The paradigm was compared to SSVEP to examine relaxation and was tested using a commercial-grade EEG, which selected for consideration of the impact of convenience on the user experience. However, significant work was needed to synchronise the data using a two-part RF device. In addition, ICA preprocessing was required to remove noises associated with poor data quality and movement. The algorithms consisted of xDAWN spatial filtering and covariance matrices, and a LR classifier was used with nested cross-validation. The statistically significant results showed that 88% of participants found the face paradigm more relaxing. This confirms the added value of convenience, which contributes to user experience and system evaluation. The classification accuracy of 67% was lower than the 80% obtained with the same set of electrodes but using a research-grade EEG. The difference is consistent with what was found in the literature.

It is important to study improving user experience because classification robustness and user convenience are equally important factors for establishing a market-ready technology that will sustain real-life use. The next chapter discusses an emotion recognition application for solving a long-standing problem of violin quality definition. The same technology could be used to further improve user experience by applying quantitative-objective measure of cognitive load.

References


[18] A. M. Dreyer, C. S. Herrmann, and J. W. Rieger, “Tradeoff between user experience and BCI classification accuracy with frequency modulated steady-


Chapter 5

A novel use of emotion BCIs: Classification of violins using EEG to determine violin type perception

5.1 Introduction

The previous chapters have provided insights on improving BCI performance with regards to classification accuracies and user experience. This chapter delivers the value of innovation by examining an application of this technology to solve a problem in the music industry. The work in this chapter has won the first prize in the School of Electrical and Electronic Engineering Postgraduate Poster Showcase. The poster is attached as Appendix B.

It is difficult to understand why music affects us emotionally. There is considerable debate whether our experiences when we listen to music are linked to past emotional experiences or if specific pieces of music have distinct acoustical parameters that evoke emotion [1], [2]. For example, happy music is often fast and written in a major mode whilst sad music is often slow and in a minor mode [3]. Psychologists have heavily researched the relationships between musical features and the physiological measures of arousal and valence. Nevertheless, many questions about the emotional responses that people have to music still remain, including those about the instruments that are used to create it. Notably, these emotional
responses may help expand our present understanding of musical instrument quality analysis.

Within the field of music perception, musical instrument quality analysis continues to develop and incorporate new technologies. It has been said that there is ‘no [objectively measurable] specification which successfully defines even coarse divisions in instrument quality’ [4]. However, many possible solutions to this problem have been proposed by researchers, including comparing the structural acoustics of different-quality violins [5]. Others have investigated the quality of famous violin makes, such as ‘Stradivarius’, by comparing these violins to new violins in double-blind listening tests [6], whilst others have studied how properties like handling experience impact quality perception [7]. A more recent study [8] conceptualised violin quality by using psycholinguistics to analyse spontaneous verbal descriptions of violin quality made by experienced musicians. A useful addition to these methods would be an automated classification of instrument quality based on the direct measurement of brain responses. This could be achieved by EEG recognition algorithms capable of reliably distinguishing between different stimuli to assess instrument quality, such as that of a violin. An earlier pilot study [9] was conducted to compare the sound of an electric violin to that of an emulated acoustic violin using EEG, and located a difference in the frontal asymmetry of the alpha brainwaves that resulted in more approach and acceptance behaviours when the better-quality violin was listened to. The frontal asymmetry index represents the valence level in the brain, and it was found to be different for the two different-quality violins. The pilot study was conducted using a commercial-grade EEG and is attached as Appendix C. Based upon measuring valence, violin-type classification test is intrinsically coupled with the assessment of emotion from music, but an investigation of this was not included in the pilot study [9]. This method could potentially provide an objective measure for musical instrument quality perception, one dependent on subconscious behaviour instead of the current self-report methods that can be unreliable for a number of reasons, e.g. see [10]–[12]. Based upon measuring valence, violin-type classification is intrinsically coupled with the assessment of emotion from music, but an investigation of this was not included in the pilot study [9].

EEG measures brainwaves from the scalp, and it has been used in many studies
related to music, such as to analyse brain behaviour during music listening and performance tasks [13]–[17] and to optimise music streaming services [18]. EEG has also been used to assess the efficacy of music therapy in treating anxiety and depression [19]–[21], and researchers have found that music shifts frontal alpha asymmetries and increases frontal mid-line theta activity in depressed individuals. Various emotion recognition models can investigate the brain behaviour of participants when they have emotional responses to music, and some models can recognise up to six emotions, namely fear, frustration, pleasure, disgust, surprise and satisfaction [22]. One model that is widely used in emotion recognition studies is the arousal-valence model [22], [23]. This model has a 2D coordinate system: the x-axis represents valence with positive values indicating happiness and negative values indicating sadness whereas the y-axis represents arousal, meaning emotional intensity. For example, high arousal could be excitement or anger whereas low arousal could be happiness or sadness. EEG is mapped onto this model by detecting fractal dimension changes in the brain [21]. Another frequently used model analyses the alpha-power (8-13 Hz) asymmetry detected from the spectral differences between symmetric electrode pairs at the anterior areas of the brain [24]. A lower alpha power in the left hemisphere than in the right is associated with feelings of joy, happiness and well-being, whereas a higher alpha power in the left hemisphere relates to negative feelings such as sadness and anger [25], [26]. It was also reported by [27] that beta frontal asymmetry is observed for happy versus sad music.

Several datasets to enable the physiological study of emotive music have been made public, and the literature analysing these datasets is often used to benchmark research in terms of data quality, algorithm development and analysis. One popular dataset is DEAP [28], which includes 32 subjects watching 40 one-minute musical video clips that were categorised according to their different arousal and valence levels. The authors’ published classification accuracy was 62% for two-class arousal states and 57.6% for two-class valence states. A study by [29] analysed emotional responses by applying dual-tree complex wavelet packet transform time-frequency features and support vector machines to DEAP, and the researchers reported accuracies of 65.3% and 66.9% for valence and arousal, respectively. Similarly, a study [30] used a Bayesian classification method and reported accuracies of 66.6% and 66.4% for valence and arousal, respectively. Many different studies
have had similar results, such as [31], [32]. A recent study that compared several datasets for accuracy — including DEAP, MAHNOB and DECAF — placed the best reported accuracies within the range of 57% to 62% [33], and our classification accuracies will be compared to those in this chapter.

In this chapter, it was aimed to design and implement a low-level EEG recognition system for violin type classification as a proof of concept that would deepen the present understanding of brain behaviour associated with musical instrument quality perception. However, given the difficulty in replicating psychological results [34], it was also sought to replicate earlier results in the literature that found distinct EEG responses to different emotional expressions in music, removing this as a confounder in the violin type analysis. Furthermore, a survey was conducted to examine the expressed emotions of the musical clips which were used in the experiment. This chapter opens by Section 5.2 describing the experiment setup and then provides an overview of the algorithms used. Afterwards, the results of the three experiments are presented and discussed in Section 5.3.

5.2 Methods

5.2.1 Introduction

This section describes the methods used to conduct the experiments. Firstly, the setup of EEG listening tests are explained in detail. Secondly, the survey that was used to confirm the expressed emotions of the clips is described. Finally, the algorithm that was used for classification is explained.

5.2.2 EEG listening tests

A professional composer was asked to produce five emotive pieces of music that express five emotions with different levels of valence and arousal: happiness (high valence, low arousal), sadness (low valence, low arousal), excitement (high valence, high arousal), threat (low valence, high arousal) and neutrality. The inclusion of different emotions ensured independence from brain behaviour related to a single emotion type, which would lead to a more generalisable and therefore stronger
classification of violin type. The musical clips were played on two different quality violins which had distinctive timbral properties. The first one was a high quality modern acoustic violin, and the second was a modern, mass-produced electric violin. Each piece was played by a professional musician several times to ensure that the different-quality violin recordings matched. These clips were then edited to ensure perfect synchronisation.

There were six sections on the experiment presented to the users as a web page, with a 30-second break at the end of each section. Each melody was played by two anonymous violins in a randomly allocated order. There were 56 clips in total, and the clips were randomly allocated to each section (i.e., a single section had random clips from the 56 clips). The melodies were trimmed into small 30-second clips. There were 16 participants in the experiment, all of whom had completed more than ten years of formal music education. The average participant age was 23.1 years, and male-female participation was equal. 56% were undergraduate students enrolled in a music program, 31% were postgraduate students, and 13% were graduate employees. All experiments occurred at 11:00 AM, apart from three experiments that were within one hour of the planned time due to time constraints. Participants were asked to refrain from consuming caffeinated drinks, such as coffee or tea, on the day of their experiment. However, no tracking of caffeine consumption took place. All experiments occurred in the same room with
the same average lighting. A 32-channel actiCHamp EEG recorder with 500Hz sampling frequency was used for the experiment. Each electrode was connected to the skin using a special conductive gel. The impedance was ensured to be less than 5 kΩ. Placing the cap and applying the electrodes to the head took 15-25 minutes on average. The time required for a single experiment was approximately 30-40 minutes. The experiment was thoroughly explained to each participant prior to starting it. The participant was explicitly asked to choose the violin they preferred more, with no mention of its actual quality. The experiment’s procedure, equipment, and recruitment and payment methods were ethically approved by both The University of Manchester, where the experiments took place, and The Royal Northern College of Music.

5.2.3 Expressed emotion survey

A survey was conducted to test the similarity between the intended emotion and the perceived emotion of listeners for the five musical clips. Thirty participants with musical experience listened to the clips that used the acoustic violin and rated their expressed valence and arousal as one of three levels: high, low or neutral. The information from this survey formed a basis for classification. Therefore, it was necessary to limit the number of options to three. High level represented 1 in classification, low level represented 0 and neutral excluded the clips from classification. A future survey could use a continuous scale for more accuracy, which is expected to yield better results [35].

5.2.4 Riemannian CSP spatial filtering and LR classification

As seen in Figure 5.2, the analysis algorithms used in this analysis have three main stages: data pre-processing, feature extraction and classification, which uses machine learning to train and test the data to obtain system accuracy. The accuracy benchmark was based off the literature, as described previously. To prepare the data, the first step was to remove the low-frequency bias present as an offset in each channel. A high-pass filter with a cut-off frequency 0.5 Hz was applied to the raw data, followed by a 50 Hz notch filter. The next step was to use an algorithm to remove electromyographic (EMG) and electrooculographic (EOG) noises, which
Figure 5.2: Overview of the algorithm. Acronyms: Independent component analysis (ICA), common spatial pattern (CSP).

are associated with muscle and eye movement, respectively. An independent component analysis (ICA) was ran to view the independent components of the data, find irregularities, and remove them.

The second step involves feature extraction and spatial filtering. The raw covariances of the trials were used as features, and this strategy was first introduced by [36]. Let $\mathbf{X}_l$ be a input data of a trial $l$ of length $T$. A sample covariance matrix (SCM) is defined as

$$SCM(\mathbf{X}_l) = \frac{1}{T-1}\mathbf{X}_l\mathbf{X}_l^T.$$ 

(5.1)
To detect features from different areas of the brain, the two intra-class covariance matrices obtained by using the two conditions (valence high vs. low, arousal high vs. low, or electric vs. acoustic violin, depending on the test) were simultaneously diagonalised. A supervised spatial filtering algorithm utilising Riemannian geometry-common spatial patterns (CSPs) was used to reduce the number of channels to 8 virtual channels. CSP was also used to view where the difference between the classes lies. The CSP graphs will show increased and decreased brain activity in the spatial patterns across the brain. In addition, the frequency bands delta (0-4 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz) and gamma (+30 Hz) were separated before running the transformation, which will indicate any brain behaviour related to each frequency band. To apply Riemannian CSP spatial filtering, the mean of a single class’s covariances was needed as described by [37]–[39]. Given symmetric positive definite (SPD) matrices $P_1, P_2, ..., P_L$ the Riemannian mean $RM$, as defined in [39], is

$$RM(P_1, P_2, ..., P_L) = \arg \min_{P \in P(n)} \sum_{i=1}^{L} \delta_R^2(P, P_i), \quad (5.2)$$

where $\delta_R(P_0, P_1)$ is the Riemannian distance between two symmetric positive definite SPD matrices, $P_0$ and $P_1$ in $P(n)$, and is given by

$$\delta_R(P_0, P_1) = \left[ \sum_{i=1}^{L} \log^2 \frac{\lambda_i}{1 - \lambda_i} \right]^{1/2}, \quad (5.3)$$

where $\lambda_i$ are the eigenvalues. Riemannian distance approximates Euclidean distance in the Euclidean feature space, which measured the separability of classes in this feature space. Next, the discriminative power of each spatial filter $w_j$ is related to its associated eigenvalue by the relationship

$$(A_0 + A_1)^{-1}A_0w_j = \lambda_jw_j, \quad (5.4)$$

where $A_0$ and $A_1$ are the Riemannian mean matrices of classes 0 and 1, respectively. Thus, the filter was chosen based on what values maximise the quantity $|\lambda_j - 0.5|$. The filtered output $\hat{X}$ therefore was obtained by the direct projection of SCM on each trial with the spatial filter $W$ that maximised the covariance be-
tween the classes:

\[ \hat{X}_l = W^T SCM(X_l). \]  

(5.5)

The final stage was to train and test the classifier for each subject using a logistic regression classifier [40]. Let \( \hat{X}_l \) be the feature vector of trial \( l \), where \( y_l \) is the label 0 or 1 indicating its class. The hypothesis or probability of the occurrence function \( h \), as defined in [41], [42], is a sigmoid with an output of range \([0, 1]\), calculated by

\[ h_\theta(\hat{X}_l) = \frac{1}{1 + \exp(-\hat{X}_l \times \theta^T)}, \]  

(5.6)

where \( \theta \) is a weight vector of same length as the feature vector \( \hat{X}_l \). The output, meaning the class label, is defined as

\[
y_l = \begin{cases} 
0 & h_\theta(\hat{X}_l) < 0.5 \\
1 & h_\theta(\hat{X}_l) \geq 0.5.
\end{cases}
\]  

(5.7)

The function is trained by minimising the cost function with the parameters of the weight vector \( \theta \).

To avoid bias, a 10-fold nested cross validation method was employed. The data were split with a ratio of 75% for training and 25% for testing. The testing trials were chosen randomly across all the trials. All accuracies were calculated using the AUC, and were verified by the t-test, which were explained in Sections 2.5.7 and 2.5.8, respectively.

5.2.5 Summary

This section described the methods used in the experiment and in the analysis. Firstly, the EEG listening tests were explained. The musical clips were composed to express 5 different emotions, which were validated by a survey. The experiment setup utilised a web-based server to allow the user to play the music in user friendly environment. Secondly, the classification algorithm was comprised of Riemannian CSPs for spatial filtering and covariance matrices as features. LR was used for classification because of its good performance with small number of trials.
Table 5.1: Expressed emotion survey results; 30 participants. Group opinion for each clip valence and arousal percentages (%)

<table>
<thead>
<tr>
<th>Intended emotion</th>
<th>Group opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High (%)</td>
</tr>
<tr>
<td>‘Happiness’ valence</td>
<td>83</td>
</tr>
<tr>
<td>‘Happiness’ arousal</td>
<td>87</td>
</tr>
<tr>
<td>‘Threat’ valence</td>
<td>0</td>
</tr>
<tr>
<td>‘Threat’ arousal</td>
<td>100</td>
</tr>
<tr>
<td>‘Neutrality’ valence</td>
<td>90</td>
</tr>
<tr>
<td>‘Neutrality’ arousal</td>
<td>10</td>
</tr>
<tr>
<td>‘Excitement’ valence</td>
<td>97</td>
</tr>
<tr>
<td>‘Excitement’ arousal</td>
<td>70</td>
</tr>
<tr>
<td>‘Sadness’ valence</td>
<td>0</td>
</tr>
<tr>
<td>‘Sadness’ arousal</td>
<td>10</td>
</tr>
</tbody>
</table>

5.3 Results and discussions

5.3.1 Emotional response survey

This survey was conducted to determine the group perception of valence and arousal for each melody. Although each melody was designed to ensure the evocation of a specific emotion, it was still important to validate and compare the intended and the delivered perceptions of valence and arousal. All 30 participants had musical experience, 16 of which were the original EEG test participants. Their input was taken after their respective EEG experiments. Table 5.1 shows the results from all participants. For most of the clips, the intended emotion matched the group opinion. However, the ‘Happiness’ clip was meant to be calming (i.e., low arousal) but the results showed it to have high arousal. The ‘Neutrality’ clip induced high valence but low-neutral arousal. A small number of participants made dissimilar choices. For this reason, it was decided to use the group’s majority input instead of personal input to define the classes. However, it is justifiable that the intended emotions and the perceived emotions will not fully match because the perception depends on personal music education, experience, and culture.

5.3.2 Replication of emotion EEG classification

To confirm the validity and rigour of the methods undertaken in this chapter, emotion classification tests completed by other researchers were replicated. In addition, EEG valence and arousal responses were examined using CSPs.
Table 5.2: Valence (high versus low) and arousal (high versus low) classification accuracies (AUC); 16 participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Valence</th>
<th>Arousal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78</td>
<td>81</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>79</td>
</tr>
<tr>
<td>3</td>
<td>81</td>
<td>73</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>77</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
<td>59</td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>72</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
<td>70</td>
</tr>
<tr>
<td>9</td>
<td>68</td>
<td>74</td>
</tr>
<tr>
<td>10</td>
<td>67</td>
<td>65</td>
</tr>
<tr>
<td>11</td>
<td>72</td>
<td>55</td>
</tr>
<tr>
<td>12</td>
<td>73</td>
<td>51</td>
</tr>
<tr>
<td>13</td>
<td>56</td>
<td>69</td>
</tr>
<tr>
<td>14</td>
<td>73</td>
<td>57</td>
</tr>
<tr>
<td>15</td>
<td>79</td>
<td>84</td>
</tr>
<tr>
<td>16</td>
<td>76</td>
<td>70</td>
</tr>
<tr>
<td>Average</td>
<td>70.71</td>
<td>69.44</td>
</tr>
<tr>
<td>P-value</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

First, training data with class definitions belonging to either low or high valence and arousal were required for the classification process. The labels that define the classes for each valence and arousal level were obtained from Table 5.1. Therefore, the group majority opinion was treated as the ground truth for the definitions of high and low valence and arousal. The analysis involved testing valence and arousal levels for each of the frequency bands separately, and this was consistent with the EEG response analysis of each band. Table 5.2 shows the classification accuracy for the group opinion taken from the survey results. The average accuracies were from 70.71% and 69.44% for valence and arousal, respectively, with p-values less than 0.0001 that makes them statistically significant. The highest valence accuracy was 81% using participant number 3, whereas the highest arousal accuracy was 84% using participant number 15. Comparing these results with the literature we find that these accuracies are higher than those obtained by [29], [30], [33] for various datasets. This indicates success of data collection and classification methods.

Second, to investigate brain behaviour in depth, CSPs were analysed for valence and arousal. Figure 5.3 shows the two CSPs with the highest energy difference. Delta frequencies indicated high frontal asymmetry for both valence and arousal. For valence, our results showed theta activity in the occipital cortex and alpha activity differences in the front and back part of the brain. In addition, beta had au-
Figure 5.3: The two maximum CSPs for valence and arousal for each frequency band. The red colour indicates increased brain activity in a certain area of the brain. Blue indicates decreased brain activity.

ditory cortex brain activity. The differences between the right and left central regions are notable for their valence levels, and these changes in brain activity were also found in musicians during both improvisation and music listening. For example, a study [43] found strong alpha and theta rhythm responses in the occipital and fronto-occipital regions of the brain during mental improvisation and listening to raga music. Further support comes from [44] who found that when participants were exposed to joyful pieces, connectivity was increased in the frontal and frontal-parietal regions. For arousal, our results showed increased brain activity in the right hemisphere of the brain and decreased activity in the left hemisphere for the delta and theta bands. Moreover, we found increased activity in the left hemisphere of brain and decreased activity in the right hemisphere for alpha bandwidths. Central regions for arousal were activated for both beta and gamma bandwidths. Different stimulation levels of the auditory cortex were apparent. These findings support the notion that change in basic musical features triggers emotional responses in listeners. A previous study has observed a significant frontal asymmetry in the frontal electrode pair FC3 and FC4 in response to pleasurable
These findings are consistent with the literature, and therefore, confirm the validity of the methods used in our research. The same methods are applied to investigate brain behaviour associated with violin type and are discussed next.

5.3.3 Violin type classification

This main focus is to test different-quality violins for classification and to examine how they affect EEG responses. Four different tests were used to obtain different information about the machine learning process. Firstly, a direct classification of violin type. This is the same test method used for emotion classification, as demonstrated earlier. Secondly, a classification that separated the five emotions to view any dependence between violin type and emotion. Thirdly, a novel classification that paired clips with the same melody but different violins together. This was accomplished by combining the EEG data of adjacent violin clips into a single feature vector, subtracting their feature vectors, and defining their training labels as 0 or 1 depending on if an acoustic or electric violin came first in the original listening test (i.e., acoustic minus electric takes 0 and electric minus acoustic takes 1). This operation removed any extra energy bands related to the melody of specific clip and kept the energy bands related to violin perception. A successful classification of this pairing process would indicate whether the violins are separable from the EEG data. Finally, an analysis of the common patterns associated with violin-type perception was completed using the CSP method described earlier.

In the first test, the classification is tested by defining violin labels as 0 for electric and 1 for acoustic. Table 5.3 shows the accuracies for all frequency bands. The accuracies were significantly lower than those obtained for valence and arousal classification. The accuracies around 50% chance level indicate the classifier failed to identify brain changes related to violin type. The p-value of 0.2225 confirms this finding. The following test expands on this observation.

The second test involved the separation of different emotions followed by classification. The data from different participants were combined to form a larger number of training and testing trials because the number of training trials per participant was now smaller. This gives a subject-independent classification, with leave-
Table 5.3: Electric versus acoustic violin classification accuracies; 16 participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Accuracy (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
</tr>
<tr>
<td>5</td>
<td>65</td>
</tr>
<tr>
<td>6</td>
<td>55</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
</tr>
<tr>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>53</td>
</tr>
<tr>
<td>12</td>
<td>35</td>
</tr>
<tr>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>14</td>
<td>31</td>
</tr>
<tr>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>16</td>
<td>46</td>
</tr>
</tbody>
</table>

Average 45.88
P-value 0.2225

Table 5.4: Emotions separated electric versus acoustic violin subject-independent classification accuracies (AUC); 16 participants (Leave-one-subject-out cross-validation)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Happiness</th>
<th>Neutrality</th>
<th>Sadness</th>
<th>Threat</th>
<th>Excitement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>63</td>
<td>44</td>
<td>45</td>
<td>52</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>56</td>
<td>47</td>
<td>59</td>
<td>56</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>56</td>
<td>53</td>
<td>63</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>76</td>
<td>50</td>
<td>58</td>
<td>35</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>38</td>
<td>58</td>
<td>55</td>
<td>52</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>63</td>
<td>53</td>
<td>49</td>
<td>40</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>50</td>
<td>54</td>
<td>49</td>
<td>43</td>
</tr>
<tr>
<td>8</td>
<td>56</td>
<td>83</td>
<td>22</td>
<td>55</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>56</td>
<td>81</td>
<td>69</td>
<td>43</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>64</td>
<td>50</td>
<td>36</td>
<td>57</td>
<td>56</td>
</tr>
<tr>
<td>11</td>
<td>57</td>
<td>56</td>
<td>31</td>
<td>41</td>
<td>80</td>
</tr>
<tr>
<td>12</td>
<td>32</td>
<td>94</td>
<td>47</td>
<td>45</td>
<td>60</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
<td>38</td>
<td>64</td>
<td>57</td>
<td>36</td>
</tr>
<tr>
<td>14</td>
<td>67</td>
<td>69</td>
<td>56</td>
<td>69</td>
<td>56</td>
</tr>
<tr>
<td>15</td>
<td>68</td>
<td>94</td>
<td>39</td>
<td>53</td>
<td>40</td>
</tr>
<tr>
<td>16</td>
<td>60</td>
<td>63</td>
<td>47</td>
<td>46</td>
<td>80</td>
</tr>
</tbody>
</table>

Average 56.01 62.63 48.66 51.41 51.46
P-value 0.0645 0.0106 0.6598 0.56 0.683

one-subject-out cross-validation. This cross-validation method iterates by training on all participants except one and testing on this participant’s data, repeated 16 times. Table 5.4 shows the accuracies for each emotion and reporting the accuracy for testing each participant. The highest accuracy of 62.62% was for the ‘Neutrality’ clip. The ‘Happiness’ clip achieved a 56.01% accuracy. The remaining emotion values were around chance-level and varied between 48% and 51%. The only statistically significant p-value that was 0.0206 associated with the ‘Neutrality’ clip. However, these accuracies are higher than those achieved by the previous violin
type classification method. As a result, the separation of emotions lead to better classification accuracies for some emotions. In addition, as the test was subject-independent, its accuracies were lower than those of the individual participants in emotion classification in Section 5.3.1, which is consistent with the literature when training and testing from different people [46].

The third test paired the electric and acoustic violins as described in the beginning of this section. Therefore, violin type classification could be tested independently from melodies. This was achieved by subtracting the power spectral density values, obtained using Welch’s method [47], in the range of 2-40 Hz of both violin clips of the same melody and then feeding the data to the classifier (see Figure 5.4). Their defined labels used for classification would then become their order (i.e., 1 for a preceding acoustic violin and 0 for a preceding electric violin). As a result, the classification of 0 or 1 indicates the violin type of the two clips. The classification testing of each participant was done separately with a nested cross-validation as defined in Section 5.2.4. The results are shown in Table 5.5. The overall accuracy is 61.56%, which is generalisable to all emotions and melodies, is higher to that in the first non-pairing test. The p-value was 0.0229, which is considered to be statistically significant. This accuracy is comparable to some classification accuracies ranging from 57% to 62% obtained from the literature as seen in [33] for various datasets. However, this accuracy is lower than those obtained in valence and arousal classification tests, which is expected from a more subtle test than valence and arousal. As a result, the EEG power spectral density features (2-
Table 5.5: Paired electric and acoustic violin classification accuracies; 16 participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Accuracy (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>2</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
</tr>
<tr>
<td>7</td>
<td>83</td>
</tr>
<tr>
<td>8</td>
<td>87</td>
</tr>
<tr>
<td>9</td>
<td>77</td>
</tr>
<tr>
<td>10</td>
<td>57</td>
</tr>
<tr>
<td>11</td>
<td>80</td>
</tr>
<tr>
<td>12</td>
<td>47</td>
</tr>
<tr>
<td>13</td>
<td>73</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>16</td>
<td>53</td>
</tr>
</tbody>
</table>

Average 61.56
P-value 0.0229

40 Hz) fed to a classifier were sufficient to determine violin type from the brain, given that the test is done on multiple violins playing the same melody.

The fourth test investigated the common patterns associated with violin type perception to view their spatial features. Figure 5.5 shows the EEG responses associated with perception recognition of the timbre. They were found by subtracting the EEGs of both violins combined from all participants during listening to the clip. The plots were separated into frequency bands to be consistent with the previous analysis. In Figure 5.5, a delta frontal asymmetry associated with violin type perception is present. In addition, auditory cortex activity is visible for theta, alpha, beta and gamma. Moreover, increased activity can be seen in the central region of the brain for beta and decreased activity in the same area for gamma. To the best of our knowledge, no previous studies have examined such brain behaviour associated with music instrument quality perception. In response, we examined the literature for studies reporting similar brain behaviour. Several studies were found that concluded music-evoked emotion processing is related to auditory cortex activity. For example, a study by [48] showed that musical expertise affects how the brain processes multi-sensory information within the auditory cortex. A study by [49] found that the auditory cortex has network nodes crucial to processing auditory information related to emotions. Another study by [50] asserted that across-timbre above-chance classification of emotions with affective vi-
Figure 5.5: The two maximum CSPs for violin types for each frequency band. The red colour indicates increased brain activity in a certain area of the brain. Blue indicates decreased brain activity.

Violin stimuli could be achieved using patterns in the bilateral auditory cortex. Our research is an opposite of that by finding an above-chance classification of timbre across different emotions. Further research could establish whether the findings of the present study relate to a positive emotional engagement with the more pleasant timbre of the acoustic compared to the electric violin, or whether our results relate primarily to auditory processing of timbral qualities. However, the classification of timbre (i.e., violin type) could be associated with violin quality. In our experiment, 15 out of 16 participants manually and anonymously rated the acoustic violin to be preferable. One participant rated the electric violin to be preferable. This could be used to future investigate the relation between timbre, quality and preference, and EEG behaviour. Nevertheless, there are limitations associated with the hypothesis that violin type classification relates to violin quality or pref-
ence. Firstly, there is no control over the background of the listener and, therefore, any intended perception of emotion or quality could be biased by music education, culture, or previous experiences. As a result, the data has uncertainty regarding the psychological perception of the stimuli. Secondly, the choice of acoustic and electric violins could confuse the listeners to relate to the recognition of real vs. unreal sounds resulting in a different perception.

Although significant developments have been made in identifying how music-evoked emotions affect the brain and, as a direct consequence, in clarifying how human emotions function [51], there is still no clear understanding of how the brain processes the emotions carried by music [50]. Thus, this experiment sheds light on the brain behaviour associated with violin type perception. This experiment is the first to use EEG and machine learning to examine violin type and its relation to brain behaviour and emotional response. The research employed EEG data and machine learning to both classify violin type and determine areas of interest in the brain.

5.4 Chapter summary

The aim of this chapter was to attempt violin type classification using an emotion recognition BCI. Three experiments were completed. Firstly, the intended and perceived valence and arousal of each audio clip was compared using a survey with 30 participants. The results confirmed the intended expressed emotions for the audio clips with the exception of ‘Happiness’. For this reason, the majority opinion was used to define valence and arousal levels. Secondly, the EEG data was analysed and resulted in classification accuracies of 70.71% and 69.44% for valence and arousal, respectively. These accuracies were consistent with those in the literature. CSPs were then analysed. For valence, theta activity differences in the occipital areas and alpha activity differences in the front and back part of the brain were found. For arousal, the frontal asymmetry of the brain for the delta and theta bands was notable. Thirdly, the accuracies of different violin type classification methods were compared. The highest reported accuracy, 61.56%, was achieved by pairing EEG data of the same melody with both violins into the same feature vector. CSPs revealed distinct activity in the auditory cortex related to vi-
olin type perception. This chapter has demonstrated a possibility for violin type classification that can be used to further examine brain behaviour related to musical instrument quality perception.

References


Chapter 6

Conclusions and future work

6.1 The purpose of the study

The scope of this thesis included visual and emotion recognitions. The main objec-
tive of this thesis was to make a significant improvement, covering many as-
pects of BCIs including robustness, human-centred development, and innovation.
Firstly, robustness was delivered by analysing and criticising the literature for ap-
propriate uses of cross-validation methods. In addition, different state-of-the-art
classifiers were tested to ensure the right classifier was chosen for the scope of this
research. Secondly, human-centred aspects such as user experience were investi-
gated. Many studies have examined classification accuracies as performance mea-
sures. In this study, user experience was also studied to improve usability and re-
duce user fatigue. This was accomplished by replacing flickering lights with pic-
tures of faces and non-faces to enable classification. However, implementation of
this paradigm created significant technical challenges and obstacles, for each chal-
lenge engineering contributions were required, such as, an electronic device to en-
able synchronisation of a low-cost EEG recorder. Thirdly, innovation was pro-
duced by applying the aforementioned values of robustness and human-centred
considerations in the creation of a novel system that classified violin types from
the brain, with the aim of providing insights on brain behaviour associated with
the quality of a musical instrument.
6.2 Thesis summary and contributions

Chapter 2 provided the EEG and mathematical background for developing BCI systems within the scope of the thesis. A BCI system requires several technical development steps to become useful. Firstly, the data is collected using EEG recorders, which, currently, are offered in either research-grade or commercial-grade qualities. A trade-off between cost and quality is needed. Secondly, the data is spatially filtered using methods like common spatial patters (CSPs) and xDAWN. CSPs are useful for extraction of frequency information, which is required for emotional analysis, and xDAWN is useful for ERP information used in visual BCIs. Thirdly, features are extracted from data in forms, such as, time-points and frequency information. Finally, the features are classified using analytical methods such as logistic regression (LR), support vector machines (SVMs), and deep neural networks (DNNs).

Chapter 3 examined classification performance by improving the reliability and robustness of different algorithms. Notably, emotion BCI accuracy expectations varied dramatically between studies, which used similar algorithms, and this study determined that different cross-validation methods significantly affected these accuracies. Some studies reported a two-class chance level above 63%, while others reported accuracies without testing their algorithms on new data. An investigation of nested and non-nested cross-validation methods was conducted by re-implementing previous studies because of these findings. It was found that the use of non-nested cross-validation methods on CNNs provided accuracies 30% higher than those of nested cross-validation methods. This can cause confusion for new researchers who attempt to benchmark their own algorithms. This study recommends that researchers clearly explain their cross-validation methods, which allows other researchers to better manage their performance expectations. Moreover, an algorithm was developed using Riemannian geometry-based CSPs, and xDAWN and was testing using three classifiers; LR, SVMs, and DNN. The three classifiers were then tested on two public datasets relevant to Chapters 4 and 5. The results showed that LR is the best performing classifier within the scope of the thesis.

Chapter 4 discussed BCI user fatigue caused by continued exposure to flickering
lights. The aim of this chapter was to improve user experience and, therefore, improve system performance. Many studies have examined classification accuracies as system performance measures, but few have examined user experience. The proposed ERP system involved the classification of face and non-face pictures instead of flickering lights, and then answer yes and no questions. To confirm the convenience of this new system, a survey of fifty people was conducted by running an SSVEP-based BCI system and the new system without recording EEG and asking the participants a series of double-blind questions about fatigue. 88% of participants found the new system to be more convenient than the SSVEP system. Moreover, the system utilised a commercial-grade EEG recorder that was quick and convenient for users. A wireless RF-based broadcaster was used because the machine required accurate synchronisation with the visual stimuli. System accuracy was tested using the same algorithm discussed in Chapter 3. The new system’s accuracy was 67%, which is consistent with accuracy expectations from current commercial-grade EEGs. The classification of face and non-face, as shown in Chapter 3, using research-grade EEG resulted in an accuracy of 94%.

Chapter 5 discussed the potential for emotion BCIs to classify different quality violins using EEG. This proof of concept could be used to determine an objective, quantifiable measure of instrument quality, which could contribute to the problem of definition of musical instrument quality. During the experiment, users listened to 56 clips randomly recorded with either an acoustic violin or unfiltered electric violin and with random emotional responses of high and low valence and arousal. The experiment was conducted using a research-grade EEG and, to reduce user fatigue and improve the results, with a user-friendly website. In the first test, low and high levels of valence and arousal were classified. This was done to confirm the competence of the system, resulting in a 70% classification accuracy. In the second test, the two violins were classified, which resulted in an accuracy of 61%. Furthermore, an objective method could be established by reverse-engineering the classification test and identifying neurological relations between quality and perception.
6.3 Future directions

Various enhancements to the thesis contributions could be done. Firstly, enhancing robustness by improving the classification performance, discussed in Chapter 3 by utilising transfer learning to solve the problem of low training data associated with deep learning. Transfer learning, which has recently shown success in improving accuracies [1], utilises deep neural network training data from previously done experiments, which are similar in the feature space, to improve model learning on new training data. Transfer learning may be useful when large amounts of labelled data were obtained from a previous task and new data are costly or difficult to obtain [1]. Thus, the goal is to utilise small amounts of new training data and adapt previous models to these new data.

The work of improving the user experience, described in Chapter 4, could be extended in various ways. Firstly, by improving the current face paradigm system performance. For example, the information transfer rate, which is proportionally related to the number of classes, could be increased by employing three-class classification instead of two-class. One way to achieve this would be by adding a category of familiar versus unfamiliar face pictures in addition to the scrambled pictures. Some previous studies have examined familiar, unfamiliar, and scrambled picture stimulations [2]. One study has investigated a stimulation with pictures of six different facial expressions: anger, disgust, fear, neutrality, sadness, and happiness [3]. Secondly, the recognition of cognitive load could be utilised to improve user experience. Design and implementation of a system that detects high cognitive load would help reduce user fatigue. For example, comparing different motor imagery tasks with respect to cognitive load to find the most personally suitable task for each user would allow customisation in the task structures and therefore improve the experience for all users. There has been interest in the field of cognitive load classification, where various studies (e.g. see [4], [5]) have managed to classify cognitive load levels from EEG. The same technology could be used to predict cognitive load associated with right hand, left hand, tongue, or foot movement, comparing their cognitive loads and give recommendations on the best suitable tasks for each user. To do this, an experiment could be setup to repeatedly perform the above mentioned tasks and then rate the cognitive loads of each task.
The work of emotion recognition of violin type, discussed in Chapter 5, could be extended in different ways. Firstly, although a proof of concept for violin classification using EEG was achieved in this study, the common spatial patterns that differentiated the violins could be further extracted to isolate specific EEG responses related to violin quality. The isolated EEG response could then be fed to a classifier and find its accuracy. This process could be repeated for all EEG responses until the specific brain behaviour is found across all subjects, which could then be utilised to quantify violin quality using subconscious behaviour. For example, the results presented in Chapter 5 indicated alpha activity in the auditory cortex was associated with violin quality perception. In this case, alpha brainwaves from the auditory cortex could be isolated and then fed to a classifier. A successful classification of this isolated EEG response indicates a correct identification of EEG response related to perception. As a result, alpha from auditory cortex could be examined in detail afterwards. Secondly, emulated violins are a popular research subject, and some studies have used double-blind listening tests to determine that there is or there is not a difference between emulated and real violins [6]. This could be confirmed by comparing electric, emulated, and acoustic violins using emotion BCIs. The test could be implemented with the same material used in this thesis. For example, filtering the existing clips recorded with electric violin to sound like an acoustic violin. However, more research into the perception of violin quality and timbre is needed. For example, in the work presented in Chapter 5, the classification of violin type (electric or acoustic) does not necessarily relate to violin quality, but rather timbre. This could cause a conflict when a listener, for example, prefers the sound of the electric violin. Therefore, for the research to correctly investigate sound quality, two acoustic or two electric violins could be compared. Moreover, the implementation of brain stimulation in improving emotion perception could potentially involve music perception. There is a number of studies examining transcranial random noise stimulation to enhance emotion perception abilities (e.g. see [7], [8]) and transcranial alternating and direct current stimulation to improve pitch memory in musicians and non-musicians (e.g. see [9], [10]). A possible setup for this experiment involves comparing musicians and non-musicians, before and after brain stimulation, and detect their abilities in perceiving sound qualities. This could enable recognising those with the highest perception ability that are superior to their peers in sound perception.
example, the listening tests comprising similar, but slightly different, quality violins and requiring the participants to differentiate the violins. The experiment to be repeated with and without brain stimulation to test the participants’ perception abilities for comparison purposes. Subsequently, utilising those with the highest perception abilities, with brain stimulation, to distinguish competitive violins and explore new acoustic features, which then could result in better understanding of quality.

6.4 Conclusion

BCI is a very young field, which means there will significant improvements happening on almost a daily basis. This thesis aimed to remind researchers of the larger picture of performance, by not only assessing accuracies in the best ways, but also including other factors, such as user experience. However, the contributions in this thesis do not only consist of criticising other’s work but also moving a step forward in innovation by using BCI technology in the music industry in such a unique way.

References


Appendices
Appendix A

SVM nested and non-nested cross-validation study

This conference paper was published in the proceedings of the European Signal Processing Conference (EUSIPCO) in 2018, held in Rome, Italy. The paper is related to the work accomplished in Chapter 3. The paper presents the effect of nested and non-nested cross-validation methods using support-vector machines. The same analysis was conducted in Chapter 3 but on a larger scale and with deep neural networks.
Abstract—Brain-Computer Interface (BCI) is a technology that utilizes brainwaves to link the brain with external machines for either medical analysis, or to improve quality of life such as control and communication for people affected with paralysis. The performance of BCI systems depends on classification accuracy, which influences the Information Transfer Rate. This motivates researchers to improve their classification accuracy as best possible. A bias problem in reporting accuracies by using non-nested cross-validation methods was thought to increase accuracy. The aim of this paper was to validate and quantify such a concept by using a low-cost commercial EEG recorder to classify visually evoking face vs scrambled pictures, and report high accuracy using non-nested cross validation. The algorithm employed Independent Component Analysis followed by feature extraction with sample covariance matrices. The data were then classified using Support Vector Machines. The accuracy was tested with nested and non-nested cross-validation methods; accuracies obtained were 63% and 76%, respectively.

I. INTRODUCTION

Brain Computer Interface (BCI) is the technology that utilizes brainwaves to link the brain with machines for various applications; including medical analysis, control of the environment, communication for those who are affected with partial or total paralysis, or any other directed purpose. BCIs use electroencephalography (EEG) as a means of measuring electrical activity in the brain via non-invasive electrodes that require no surgery or long preparation [1]. The most common BCI modalities are: Motor Imagery, P300-oddball, and Steady State Visually Evoked Potentials (SSVEP). For these systems to work efficiently, classification accuracies, often reported in Area Under Curve (AUC) percentages, must be high enough to maintain an acceptable level of Information Transfer Rate (ITR). This led to great motivation to make every effort to investigate accuracy improvements. The accuracy of visually evoked BCIs hugely depends on several factors such as the quality of the EEG recorder, experimental setup, nature of stimuli, and algorithm development.

The use of research-grade EEG recorders enhances the accuracy significantly, compared to commercially inexpensive EEG recorders due to differences in signal-to-noise ratios (SNRs). One study tested the classification accuracy using Emotiv EPOC (commercial EEG) with 14 electrodes, and a Biosemi headset (research-grade) with 32 electrodes employing the oddball paradigm [2]. They have found the accuracy of the 32-channel Biosemi headset to be 88.5% and the Emotiv to be 61.7%. Many other studies have examined performance of commercial EEG recorders employing other visually evoked BCIs like SSVEP and obtained similar results [3], [4].

A second important aspect is the type of visual stimuli and area affected by different classes stimuli. For example, many accuracies above 95% have been reported for the P300-oddball and SSVEP paradigms [5], [3], [4]. Meanwhile, face recognition based classification accuracies are rather inferior to this, just as discussed in the following.

Another important aspect is the classification algorithm, which is a multistage problem. Taking a Magnetoencephalography (MEG) classification competition of face vs scrambled images as a benchmark, which is a similar technology to EEG, the three best classification accuracies reported for subject-independent classification were 75%, 73%, and 71%, respectively. The winner also reported a subject-dependent classification accuracy of 86%. They utilized Event-Related Potential (ERPs) sample covariance matrices as features, and vectorization using tangent space along with a logistic regression (LR) classifier. The second place work involved down-sampled filtered raw-data as features with LR and random-forest (RF) classifiers combined. Whereas the third place used Support Vector Machines (SVMs). The leader-board is available via [6]. A more recent study [7] has reported the use of non-linear SVMs and an RF classifier with XsDAWN spatial filtering and reported a 71% accuracy based on EEG and 82% using MEG.

In this paper, we would like to shed light on another problem that affects reported accuracies, which is the use of nested vs non-nested cross-validation methods. This topic has generally been explained in [8]. The non-nested cross-validation method divides the data into training and testing parts, while nested cross-validation divides the data into training, validation and testing parts, forming two cross validation steps. The training data allow the classifier to learn the parameters and tune them.
for testing the validation data, without access to test data. Non-nested cross-validation makes use of testing data for validation stage and report best accuracies, thereby increasing overall accuracy.

The problem lies in the fact that it is difficult to know which method was used unless clearly indicated creating an unjustified gap between high accuracies (above 95%) and medium accuracies (70-80%). In this study we will test the possibility of obtaining robust accuracies using a low-cost EEG recorder and visual perception. This will be compared using nested and non-nested cross-validation methods, which we hope will indicate that not all high accuracies reported using low-cost recorders are actually accurate and feasible for a real-time application. This will be achieved by using the best algorithm methods from the literature and assessing them with a dataset that was made public, then it will be tested on our data using the aforementioned cross-validation methods in the hopes it will enlighten the processes used by other researchers and motivate them to clearly indicate their reporting methodologies. We will also show pseudo-codes to indicate how both methods are utilized based on the Python platform.

This is not a new problem. It could be considered a fact that in machine learning, reporting training accuracy (non-nested cross-validation) results in higher but less generalizable accuracies than nested cross-validation. However, in BCI application we believe this is still used. Otherwise, the existence of large gaps in reporting accuracies cannot be justified. The aim here is to quantify this problem in practice using SVMs in a visual BCI application and motivate researchers to progress with it.

The overall methodology used in this paper was to design an ERP recognition system in Python and collect data using a synchronized Emotiv EPOC+. That will be followed by explaining the classification algorithm consisting of preprocessing, feature extraction and cross-validated classification using SVMs. The results will include analysis of a dataset of faces vs. non-faces, using different set-ups for comparison purposes. It will then include assessing the data collected employing the commercial EEG.

II. Methodology

Participants were requested to answer a list of questions by staring at pictures. Their visual perception determined their EEG behavior and this was utilized to enable communication. The software posed each question by presenting a message window that asked simple yes/no questions. The answer screen would appear with two options Yes (left) and No (right). The participant answered the question by staring at the cross sign beneath the words Yes and No accordingly, see Figure 1a. A timer represented by a red growing bar at the top of the screen indicated when the pictures would appear. At the end of the timer, two random images appeared where the cross signs were located. One showed a face and the other was a scrambled picture. Both were randomly chosen by the software; further, the association of a face with a yes/no was randomized so the participant could not predict the picture presented. The pictures were present for 500 ms and then were replaced by pictures of circles, see Figure 1b. The database used for the pictures was obtained from the dataset published by [9]. There were 100 questions in total for each subject.

![The software screenshot prior to stimulus](image1.png)

(a) The software screenshot prior to stimulus

![The software screenshot during stimulus](image2.png)

(b) The software screenshot during stimulus

Fig. 1: The software developed in Python to enable communication using visual perception

The device used in this experiment was an Emotiv EPOC+ with a sampling frequency of 128 Hz. Ten volunteers participated. Experiments last from 30 to 45 minutes. Electrode locations (using the 10-20 electrode location system) were AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4 with references at P3 and P4 (CMS and DRL respectively). Saline was employed to wet all electrodes. For reference electrodes, smell- and color-free water-soluble based gel was also used to ensure conductivity if the saline dried out. The setup time took less than five minutes for each participant. A synchronization circuit was needed to allow precise timings to the Emotiv apparatus with channels T7 and T8 thats sends triggers when the photos appeared on the screen. A battery-based system was developed to provide a direct link to the electrodes from the stimulus software, developed with Python. Similar work has been reported by [10]. The software sent a command using serial communication to a USB-connected microcontroller (via a photodiode attached to the monitor) to transmit a radio-frequency (RF) signal to the receiver attached to the EEG recording machine. The receiving device which
was enabled with RF communication triggered one of the electrodes of the EEG machine (T8) with a number of pulses of 330 \( \mu V \) for 8 ms, equivalent to 1 sample with a sample frequency of 128 Hz. The ground of the circuit was connected to another electrode (T7). Both electrodes were biased to the DRL electrode using 500 \( k\Omega \) resistors. In the case of a failure, the resistors would limit any current flow to 9 \( \mu A \). This limit was less than 10 \( \mu A \) for the CF Applied Part according to the IEC 60601-1 requirements. The overall system is presented in Figure 2.

![Figure 2: The setup for the system. The EEG recording machine along with the synchronization circuit inside a three-dimensional (3D) printed enclosure. The thickness of the walls is less than 1 mm making the total weight low and not affecting the balance of the headset.](image)

### III. Algorithm

The classification algorithm consisted of three main components; filtering, feature extraction and classification which includes training and testing.

Raw data was fed into a 3rd order Butterworth band pass filter of cutoff frequencies between 2 and 20 Hz. This is required to remove low frequencies such as offsets and undesired high frequencies, like mains at 50 Hz. At this stage plotting the averages of the trials data in most cases did not reveal any ERP components, such as P300 or N170. More advanced filtering was needed. Independent Components Analysis (ICA) was then applied compute 12 independent components corresponding to the 12 recording channels. Independent components associated with visible ERPs were kept, either at the front with positive potential (given the reference point was at P3) and negative potentials at the occipital part the brain. The remaining components were removed. The process was carried out manually for all subjects. The noise was successfully removed and ERPs could now be seen (see Figure 3).

The ERP covariances of the data were then calculated. This resulted in a feature matrix of size \( E \times E \) for each trial, where \( E = 12 \) is the number of electrodes. The covariances, as discussed in [11], were calculated by firstly concatenating each trial \( z_i \) with the averages of each class \( p(1) \) and \( p(2) \):

\[
\tilde{z}_i = \begin{bmatrix} p(1) \\ p(2) \\ z_i \end{bmatrix}
\]

The spatial covariance matrix \( \sigma_i \in \mathbb R^{12 \times 12} \) was therefore defined as:

\[
\sigma_i = \frac{1}{N} \tilde{z}_i \tilde{z}_i^T
\]  

(1)

However, these features are in matrix form and need to be converted into vector form. To do this we needed to use Riemannian Geometry, which is explained in [12]. The Riemannian distance for two covariance matrices \( \sigma_1 \) and \( \sigma_2 \), representing classes 1 and 2, is defined by [13] as:

\[
\delta(\sigma_1, \sigma_2) = \| \log(\sigma_1^{-1/2} \sigma_2 \sigma_1^{-1/2}) \|_F = \sum_{c=1}^E \log^2 \lambda_c \]  

(2)

where \( \lambda_c \), \( c = 1...E \) are the real eigenvalues of \( \sigma_1^{-1/2} \sigma_2 \sigma_1^{-1/2} \) and \( E = 12 \) is the number of electrodes. Thus the Riemannian mean of the \( I \) covariance matrices is the matrix minimizing the sum of the squared Riemannian distances defined in [14] as:

\[
\arg \min_{\sigma} \sum_{i=1}^I \delta_2^2(\sigma, \sigma_i)
\]  

(3)

to feed the features to the classifier it is necessary to project matrices in a vector Euclidean space, referred to as tangent space, leading a covariance matrix of size \( E \times E \) to be represented by vectors of dimension \( E(E+1)/2 \). In this case \( E \) was 12.
an SVM classifier was used. SVMs were explained thoroughly in [15]. Using a Radial Basis Function (RBF) kernel, parameters conventionally known as $C$ and $\gamma$ needed tuning. Parameter $C$ represents the cost i.e. the classification surface smoothness to compromise misclassification of training trials to gain a simpler decision surface, where $\gamma$ represents the impact of individual samples on choosing support vectors. Tuning could be carried out in two ways that result in major differences in reporting accuracies; nested and non-nested cross-validation methods.

In nested cross-validation, the parameters are tuned using the training data without access to the testing data. Unlike non-nested cross-validation where the test data is used to optimize the parameters, and report scores based on best accuracies. We tested both methods and reported accuracies for comparison purposes. The values of parameters of $C$ ranged from 1 to 1000, and $\gamma$ ranged from 0.0001 to 0.1, with 10 folds each. The reported accuracy metric was in the Area Under the Curve (AUC), which is a well-known technique and more information on it could be found easily.

The pseudo code in Algorithm 1 describes one way for reporting cross-validation methods. Lines 2-5 indicate preparing the data for classification. Line 6 implements a grid-search method (based on Python) using 10-fold shuffle split of SVM. Line 7 signifies the training of the machines using the gridsearch method. The difference lies in Line 8 where the non-nested best accuracy of the grid search is reported as system accuracy, whereas the nested methodology has an extra layer of cross-validation method using 10 folds and reports average accuracy.

Algorithm 1 Nested vs non-nested cross-validation methods

1: procedure GETACCURACY
2: $cv \leftarrow$ ShuffleSplit(n_splits=10)
3: $X \leftarrow$ data (trials,channels,samples)
4: $y \leftarrow$ labels
5: $X_{\text{features}} \leftarrow$ calculate features of $X$
6: $model \leftarrow$ GridSearchCV(estimator=SVM, cv)
7: $model.fit(X,y)$
8: $non\_nested\_score \leftarrow model.best\_score$
9: $nested\_score \leftarrow$ cross_val_score(model, $X,y, cv$)

IV. RESULTS

A. Research-grade dataset validation

To test the algorithm and better analyze the original data, nested cross-validation SVMs were applied to an external dataset. This dataset was obtained from the OpenfMRI database. Its accession number is ds000117. The dataset included 16 subjects and used a total of 74 EEG electrodes and 306 MEG electrodes [9]. To properly confirm the algorithm, the same 12 electrodes were also tested in the analysis using nested cross-validation.

Table I shows the different accuracies obtained using different setup electrodes. It shows that the best accuracies of 86% and 85% were obtained using the 74 EEG electrodes and 306 MEG electrodes, respectively. Another column was added to compare the same 12 electrodes used in the Emotiv EPOC+.

The test was also conducted using different number of training trials to test the effect of having different experiment lengths; 100 (similar to the EPOC+ experiment), 200, and 300 to find the point of accuracy convergence. The results demonstrated that using 300 training trials the accuracy was 77%. Further, using 200 training trials, the accuracy was 75% and using 100 training trials the accuracy was 70%. These results were used as a benchmark for our system accuracy.

Table II lists the accuracies obtained by running the algorithm for both the nested and non-nested cross-validation methods. Non-nested cross validation suggests a large superiority to nested cross-validation. The average accuracy for non-nested cross-validation was 76% and the average for nested cross-validation was 63%. The greatest difference was for subjects 6 and 9 at an increase of 19%. The lowest increase was for subject 8 at 6%.

V. DISCUSSION

The algorithm accuracies for the dataset were similar to those obtained by the winner at the competition for the same dataset. Combining SVM with covariances resulted in similar accuracies garnered by an LR classifier. Over-fitting of the RBF kernel, owing to limitations in the number of trials and less generalizable conditions, might become an advantage if reporting non-nested cross-validation accuracies. However, analysis of the non-nested classification of the research-grade dataset was not included in this paper based on space limitations. The number of electrodes affected the accuracy;
306 MEG and 74 electrodes resulted in similar accuracies. Reducing the number of electrodes to 12 diminished the accuracy from 86% to 77%. This shows the benefit of machine learning where more data, with localization and spatial filtering, improves accuracy. In addition, the effect of the number of training trials is significant. However, high cost is associated with having large number of trials, including subjects and time.

The accuracy of the system is not as high as 95% as reported by studies using P300 and SSVEP paradigms. This could be caused by a number of factors such as the ERP pattern differentiation between the target and non-target classes. The face and non-face ERP patterns are very similar to each other when looking at N170 and P300 components. However, for the oddball paradigm, the ERP patterns of the target class have significant P300 components while there are no ERP components in the non-target class. The same applies to SSVEP where the frequencies are different for various stimuli classes.

Accuracy obtained by research-grade EEG recorders is expected to be superior to commercial EEG recorders. The algorithm accuracy when tested on the dataset with the research-grade EEG recorder resulted in an accuracy of 86% using a nested cross-validation method. This, however, was obtained by using 300 trials for training and 74 EEG electrodes. With the same 12 electrodes, as in the Emotiv EPOC+ and a similar number for training trials at 100 led to an accuracy of 70%, which is 7% more than the data obtained by the Emotiv EEG recorder of 63%. This difference is justified by the difference of SNR between the recorders. However, low-cost easy-to-use EEG is also favorable when time funding resources are limited.

On the other hand, the non-nested classification accuracy 76% is higher than expected when compared with the same number of training trials and electrodes using the research-grade EEG. This indicates that it is possible to report high accuracies with low-cost hardware and setup. The big difference in accuracies of the utilization of nested vs non-nested is an reflection of the importance of assessment tools and their implication in real-life application, where access to validation data is limited to validation and cannot be used for reporting accuracies.

VI. CONCLUSION

This paper analyzed different factors that affect accuracy performances in visually evoked BCI systems. It focused on the aspect of reporting accuracies by using nested vs. non-nested cross-validation methods. To analyze the matter in more detail, we collected face recognition EEG data associated with looking at pictures of faces and scrambled images with a synchronized commercial EEG recorder. It showed the possibility of reporting relatively high accuracies with such low-cost equipment by simply changing the reporting methodology. The algorithm for classifying faces and non-faces was designed based on the state-of-the-art in face recognition. The algorithm was tested on a dataset and deemed acceptable. The accuracy of the system using the collected data had accuracies of 63% and 76% using nested and non-nested cross-validation methods, respectively. The nested cross-validation accuracy of 63% was compared to obtaining 12 EEG electrodes from a research-grade EEG that resulted in 70% accuracy. The non-nested accuracy of 76% was thought to be higher that the usual, which was accomplished by reporting bias. This raised the question of whether all reported accuracies in the literature obtained by either low-cost or research-grade EEG that were of very high accuracy, were accurate representations of reality. This paper hoped to encourage researchers to clearly indicate their cross-validation methodology to reduce confusion caused by reading different accuracies in the literature.

ACKNOWLEDGMENT

This work was part of a PhD research sponsored by King Abdulaziz University, Jeddah, Saudi Arabia.

REFERENCES

Appendix B

Postgraduate poster showcase winner

This poster has won the first prize in the School of Electrical and Electronic Engineering postgraduate research poster conference in November 2016. The poster introduces the concept of violin type classification using EEG. It also presents the methods and the results of the pilot study conducted prior than the work in Chapter 5. The complete results are discussed in the conference paper enclosed as Appendix C.
Emotion recognition: Quality of musical instruments based on listeners’ brain behaviour

1. Motivation

There have been many studies on what separates good and bad violins attempting to define musical instrument quality. Psychoacoustic tests are typically used to recognise a listener’s preference to violins such as double-blind listening tests\(^1\). Others compared the structural acoustics of known good and bad violins\(^2\). The aim of this study is to explore brain behaviour analysis as an alternative method for gauging a listener’s preference to a particular instrument.

Electroencephalography (EEG) is the technology used to measure brainwaves, mainly, to diagnose brain disorders such as epilepsy. It is applied by connecting a group of non-invasive sensors on different areas of the head.

2. Background

- Literature suggests emotive states could be recognised using EEG alpha signals in the frequency range 8-13Hz\(^3\). More alpha activity in the left hemisphere of the brain corresponds to more positive (happy) music and more alpha activity in the right hemisphere corresponds to more negative (sad) music\(^4\).

3. Experimental Setup

To establish EEG as a valid method of measuring preference, 2 violins of significantly different quality were used. The first violin was an emulated Stradivarius from 1732. The second violin was an unprocessed electric violin. Each violin had five 11 second samples for use in the experiment. The corresponding samples for each violin were grouped anonymously in pairs and arranged in a 2x5 grid on a user interface, as shown in Figure 3.1. Participants were asked to write down their preferences while their EEG data is recorded throughout the experiment. The setup is summarised as follows:
- The EEG machine used was a 14-channel Emotiv EPOC+.
- 4 left side electrode locations were AF3, F3, F7, FC5 (Figure 3.2).
- 4 right side electrode location were AF4, F4, F8, FC6.
- The reference electrodes were located in P3 and P4.
- The setup took less than 5 minutes on average.
- Saline was used to wet the electrodes for conductivity.
- Each participant was given a pair of earphones for music listening.
- There were 5 participants for the experiment.

4. Algorithm

In order to compare emotional responses in different violins we need to find a measure of left/right alpha asymmetry. This is called the Frontal Asymmetry Index (FAI) and is defined in Figure 5.1. \(P_x\) is the average power spectral density for side X.

The first step in calculating the power is to divide data into 2s epochs, with 0.5s overlapping. The power spectral density is then found for each epoch separately and are all averaged for each electrode. Then averaging powers across 4 electrodes for left/right. The asymmetry is then computed using the aforementioned equation. The algorithm takes the data associated with the last click of each clip.

5. Results

Figure 5.2 shows all 5 participants matched a preference for at least 3 out of 5 Stradivarius samples. One of the participants had a preference for all 5 Stradivarius samples and 2 participants preferred the Stradivarius 4 for the samples.

6. Conclusion

Although the results do not statistically prove significance of the experimental procedure, they can be used as a prelude to more sophisticated tests that provides a deep insight into quality of instruments. The next stage of the research is to expand the tests to 20 participants, with 32-ch EEG and different musical stimuli.

7. References

Appendix C

Violin type commercial-grade

EEG pilot study

This conference paper was published in the proceedings of the Institute of Acoustics 2016 conference, held in Warwick, England. The paper is related to the work discussed in Chapter 5. The paper presents preliminary results on analysing emulated and electric violins from EEG by comparing brain behaviours. The experiments conducted in Chapter 5 were on a larger scale and investigated the use of machine learning.
A PRELIMINARY INVESTIGATION INTO EMOTIONAL RESPONSES TO EMULATED VIOLINS USING ELECTROENCEPHALOGRAPHY

M Abdulaal School of Electrical and Electronic Engineering, University of Manchester, UK
T Lloyd School of Electrical and Electronic Engineering, University of Manchester, UK
P Gaydecki School of Electrical and Electronic Engineering, University of Manchester, UK

1 INTRODUCTION

There have been significant investigations into what separates “good” and “bad” violins with an extensive amount of literature attempting to distinguish between them. Psychoacoustic listening tests are typically used in the field of violin research to gauge a participant’s preference to different violin sounds. The aim of this paper is to explore and document the use of electroencephalographic (EEG) analysis as an alternative method for gauging a listener’s preference to a particular instrument.

In Fritz’s 2012 publication on old versus new violins there is speculation as to whether superficial attributes (such as price or age) may contribute to a violinist’s preference [1]. This double-blind study employed measures (for example welders goggles to obscure the fine details) to ensure that the participants could not let these ‘superficial’ qualities influence their judgement [1]. One of the aims of this study is to show that by using EEG analysis, a participant’s “true” preference can be determined irrespective of superficial qualities.

1.1 Existing Psychoacoustic studies

Woodhouse describes a type of psychoacoustic test where a “single variable” in the virtual violin’s frequency response is adjusted until the change is detectable by the listener [2]. These changes are known as “Just noticeable differences” and such tests are widely used in the field of audiology [3]. An early publication by Fritz and Woodhouse et al used this methodology to explore a listener’s perception to subtle changes in key parts of the violin’s frequency response [4]. The participants were presented violin samples in groups of 3, where one had undergone spectral adjustment involving amplitude or frequency shifting in one of the designated modes/bands. The participant would then be asked to select which violin was different. The amount of adjustment would increase until the participant could confidently identify which of the three samples were different [4].

A similar experiment by Fritz aimed to link “English Language Descriptors” with spectral changes in virtual violins [5]. The spectral response was divided into 5 frequency bands each one octave wide within the range 190-6080Hz [5]. To generate the different violins one of the bands would be amplified within ±10dB. The study presented musically trained participants with pairs of these virtual violins and asked which one was more/less X than the other, where X was one of the 5 chosen descriptors [5].

Other forms of listening tests do not follow the discussed methodology of exploring just-noticeable-differences, but instead compared different violins as whole. One of Fritz’s more recent studies aimed to compare “Old” and “New” instruments, some of which were from “classic” Italian makers [1]. The study had 21 “experienced violinists” playing the instruments and giving appropriate feedback in a hotel room, which was chosen for its “dry acoustics” [1]. The initial part of the study had the participants comparing 9 pairs of old and new violins and asked them to state which they preferred. The next part of the experiment involved the same participants evaluating the violins and selecting the “Best” and “Worst” for the categories “range of tone colours”, “projection”, “playability”, “response” which were left “undefined” [1]. The participants were also asked to select which violin they would “Most like to take home”. This study was later repeated with key changes to the experimental protocol (such as number of instruments, venues used) [6]. The experiment was
carried out in two separate locations that are typically used by violinists, the first being a practice room and the second being a 300 seat concert hall [6]. The first part of the test was repeated and instead of comparing pairs of violins the participants were instructed to select their four favourites and to “reject” any that they didn’t like [6]. The violin selected as their “favourite” along with two other violins (their own instrument and their favourite of the opposite category) were then carried forward into the next part of the experiment [6]. Following the original study the next part of the study had participants rating these three instruments on a scale of 0-10 for a range of categories which were again left “undefined” [6]. These two experiments were carried out in both venues, although for the second experiment the participants were provided with further options (for example another person playing the instrument) to help in their evaluations [6].

This brief review of existing test protocols reveals that significant care needs to be deployed in the development of a study to gauge a listener’s perception to a particular instrument or sound. The study discussed here is largely influenced by Fritz’s recent studies comparing old and new instruments, as opposed to determining the JND. However for this study the comparison is between an unprocessed electric violin and an emulated Stradivarius, rather than two “real” instruments.

2 OVERVIEW OF VIOLIN CHARACTERISATION AND EMULATION

This experiment utilises emulated (“virtual”) violins, which are widely used in perceptual listening tests carried out by other researchers [1]–[6]. Fritz and Woodhouse both state that one of the advantages of using virtual violins is that changes can be made to the response without having modify the physical properties of the original instrument [2], [4].

The emulation process begins with the characterisation of a real instrument resulting in a digital representation as a finite impulse response (FIR) filter [4]. There are several different methods of achieving this characterisation each with their relevant merits and demerits [4], [7]–[9]. For this particular investigation the characterisation was acquired by measuring the sound radiated from the violin after it has been provided with an impulse. An impact hammer is used to provide the impulse to the instrument by striking it on the bridge [7]. Spectral analysis of the impact hammer however reveals that it does not provide a perfect impulse, and thus further processing is required to compensate for this. Deconvolution of the recorded signals is carried out to provide this compensation resulting in a close approximation of the “true” impulse response of the instrument [10]. Figure 1 shows an example of a violin that has undergone this correction process.

![Uncorrected Violin Spectral Response Y(\omega)](image)

- Magnitude (dB)
- Frequency (Hz)

![Corrected Violin Spectral Response H(\omega)](image)

- Magnitude (dB)
- Frequency (Hz)

Figure 1. Spectrum of a characterized violin before (top) and after (bottom) the correction process.
The final stage is to use this digital characterisation to generate the final emulated output. This process simply involves convolving the violin’s FIR representation with the raw signal from an electric violin (or a piezo electric sensor mounted on a violin bridge) [2], [3], [5], [7]. The convolution kernel is typically executed offline using MATLAB although systems do exist to carry out this process in real time [11]. Woodhouse explains that by using a pre-recorded signal from the violin bridge, the variation in playing does not influence the listener’s perception [2].

3 MUSIC AND EMOTIONS

3.1 Emotion Detection

3.1.1 Electroencephalography

Electroencephalography is a widespread technology that measures brain activity, and is mainly used to assess epilepsy and other brain disorders [12]. The clinical devices used are highly accurate to ensure appropriate assessment. Such devices require special training for practitioners to prepare patients for the tests. Some EEG tests and readings are taken when the patient is asleep [13]. The output from an EEG device is a waveform containing significant spectral data. Often the frequency of the measured signal can be used to indicate activity state of healthy person [14]. Disorders are often identified by unusual activity in the brain. Due to the low amplitude of the recorded signals (in the μV range), it makes it a significant challenge to design an EEG device. EEG signals are often contaminated with other signals such as Electromyography (EMG) signals caused by muscular movement, ambient interference and mains noise, all of which may be significantly greater in amplitude than the EEG. This in turn requires appropriate filtering. EEG signals that are useful to interpret are in a frequency range of 2-40 Hz.

3.1.2 EEG Music Analysis

Emotions may be recognised using various brain-computer interface (BCI) techniques. That is to detect if a person is happy or sad. Some models are developed to recognise up to six emotions such as fear, frustration, pleasure, disgust, surprise and satisfaction [15]. In recent research, emotion recognition has become vital in many applications. It can be used as an assessment means in marketing campaigns, in which BCI is used with shoppers that look at a specific marketing campaign, such as an advertisement, and their emotions are recorded to test if the desired outcome is achieved. This could save marketing companies by obviating shopper surveys, and in fact getting what they feel instead of what they say they feel [16]. One of the models that is widely used in emotion recognition is the Arousal-Valence model [15], [16]. This model has a 2D coordinate system, in which the fractal dimension changes in the brain are recorded and mapped. Those changes represent the feeling towards an exciting fact, at the moment of the recording, which are by definition the emotions. Another model that is frequently used is the alpha-power (8-13Hz) asymmetry detected from the spectral differences between symmetric electrode pairs at the anterior areas of the brain [17].

Music plays a big role in affecting listeners’ emotion state and mood. Researchers found sad music was related to changes in skin conductance levels, affecting heart rate, blood pressure and skin temperature. Happy music produced changes in respiration activity [18].

Evaluation of music stimuli using EEG by measuring the alpha-power asymmetry has been achieved by many researchers to observe the brain behaviour when listening to happy music as opposed to sad or “negative” music [19]. Lower alpha power in the left hemisphere than the right is believed to be associated with joy, happiness and a feeling of well-being [19]. Conversely, higher left hemisphere alpha power (i.e. less activity) relates to negative feelings such as sadness and anger. Researchers in [19] asked participants to listen to certain musical excerpts and then express
their feeling when listening to those clips in a questionnaire. The average results for positive music showed larger differences than the negative music.

In this paper we evaluate and compare different “violins” playing the same melodies in attempt to explore whether EEG could be used in quantifying musical instrument preference.

4 LISTENING STUDY

4.1 Paradigm

4.1.1 Violin Samples

To establish EEG as a valid method of measuring preference towards a violin sound it was decided to select one “good” violin and one “bad” violin so that the participants were listening to samples that were significantly different in quality. The first violin was a “Tom Taylor” Stradivarius from 1732 which was characterised in an anechoic chamber to negate the influence of the room acoustics. The second violin was an unprocessed electric, which of course did not require any characterisation as the test just employed raw electric signal from the instrument. This recorded signal was convolved with the characterisation of the Stradivarius to provide the emulation for the “Good” violin.

Each virtual violin was 55 seconds in length and further divided into five 11 second samples for use in the experiment. The corresponding samples for each violin were grouped in pairs and arranged in a 2x5 grid on a user interface. The user interface had a grid of buttons that would play the corresponding sample when pressed as shown in Figure 2.

4.1.2 EEG Setup

The EEG machine used for the experiment was an Emotiv EPOC+ with 14 electrodes. However, only 8 electrodes with 2 reference electrodes were used for recording. The 8 electrodes allocations using the 10-20 system are AF3, F3, F7, FC5 on the left side of the brain and AF4, F4, F8, FC6 on the right side of the brain. The reference electrodes were located in P3 and P4. The setup of the experiment took less than 5 minutes on average. Saline was used to wet the electrodes to make them conductive. Each participant was given a pair of earphones for listening to the clips.

Figure 2. Experiment Paradigm: The user has control over which clip to listen to. The types of violins were unknown to the user and are randomly ordered. The user could listen to any clip as many times as desired.
4.1.3 Test Procedure

The experiment consisted of 5 participants (of which 4 were female) who listened to the different samples while being monitored by the EEG device. The participants were instructed that they could listen to the samples as many times as they wanted and in any order before selecting which violin they "preferred" for each of the 5 pairs. The idea behind this was to synchronise their selections with the recorded EEG data. One of the researchers sat next to the participants during the experiment to ensure that the EEG device was correctly recording data.

4.2 Analysis Algorithm

![Flowchart of the analysis algorithm](image)

**Figure 3.** Overview of the algorithm used to analyse the emotion state for each participant

In order to obtain and compare the emotional states between different clips we need to find a measure of left to right alpha asymmetry. We define this measure of Frontal Asymmetry Index (FAI) as:

$$\text{FAI} = \frac{\log P_{\text{left}} / \log P_{\text{right}}}{\log P_{\text{total}}}$$  \hspace{1cm} (1)

Where $P$ is the average power spectral density in the range of alpha activity between 8-13 Hz. The first step in calculating the power is to separate each data segment into ten 1.5-second epochs, each of which overlaps its predecessor by 0.25 seconds. The power spectral density is then found for each epoch separately and all are averaged for each electrode. The following step involves averaging powers across 4 electrodes for each side of the brain for each epoch. The FAI is then found using the equation (1). The algorithm takes the data associated with the last click of each clip.
## 4.3 Results and Discussion

Table 1. Written preferences of each of the participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
</tr>
<tr>
<td>2</td>
<td>Electric</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Unsure</td>
<td>Stradivarius</td>
</tr>
<tr>
<td>3</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
</tr>
<tr>
<td>4</td>
<td>Electric</td>
<td>Stradivarius</td>
<td>Unsure</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
</tr>
<tr>
<td>5</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
<td>Stradivarius</td>
</tr>
</tbody>
</table>

Table 1 shows the recorded preference for each of the participants. Most of inputs preferred the Stradivarius, 2 people reported they were not sure which instrument was better for Sample 3. Also the same 2 participants preferred the electric violin for Sample 1. The FAI results using alpha-asymmetry for each of the clips for all participants are shown in Figure 4.

![Figure 4](image-url)  
**Figure 4.** The FAI for each participant for all clips are shown above. FAI is found using equation (1). Noting that FAI is a relative measure between power in left and right brain hemispheres and is not an absolute indicator for emotions. To get that a baseline needs to be measure at the beginning of the experiment.

All five of the participants showed a preference for at least three out of the five Stradivarius samples. One of the participants had a preference for all five Stradivarius samples and two participants preferred the Stradivarius for four of the samples. The two participants that were "unsure" on a preference for Sample 3 had an increased FAI for the electric violin. Interestingly the same two participants that stated they preferred the electric violin for Sample 1 had an increased FAI for the Stradivarius samples. In total there were nineteen total preferences for Stradivarius violins detected by EEG analysis. This is an equivalent of 76% for all clips across all participants. It is noted from the results that different melodies produce different FAI. As a result, to be able to compare two different instruments the same musical phrase has to be played using the compared instruments. Considering musical phrases can be generalised as "happy" and "sad", when evaluating an instrument playing a "sad" piece a larger preference may be determined by recording more sad emotions. Similarly playing a "happy" piece may invoke more happy emotions in the brain.
Proceedings of the Institute of Acoustics

Figure 5. FAI is shown for one common clip that was clicked the most by all participants. Noting that participant 3 listened for the clip played by the Electric violin only once.

Figure 5 shows that one of the musical phrases was played more than once by most participants. Apart from Participant 3 (who only clicked the electric violin once), all other 4 participants show that they had a similar or a reversed emotion state for the two instruments. This could either justify that their choice was based on their initial hearing or that the fact that brains actually learn by repeated hearing even though two of them did not write down the change of state.

5 CONCLUSION

Although these results do not statistically prove significance of the experimental procedure, they can be used as a prelude to more sophisticated psychoacoustic tests that may eventually be considered as a new standard for quantifying different musical instruments.

6 ACKNOWLEDGEMENTS

The authors would like to thank Hans Johannsson for providing the recorded violin signal and the characterisation of the Stradivarius violin.

7 REFERENCES


